

ALAGAPPA UNIVERSITY



[Accredited with 'A+' Grade by NAAC (CGPA:3.64) in the Third Cycle
and Graded as Category–I University by MHRD-UGC]

(A State University Established by the Government of Tamil Nadu)

KARAIKUDI – 630 003

Directorate of Distance Education

M.A. (Economics)

IV - Semester

362 42

ECONOMETRIC METHODS

Authors:

Dr. Suman Lata, Lecturer, Deptt. of Economics, GDMG (PG) College, Modinagar
Units: (1, 3.5, 4.0-4.2, 4.6, 5.3, 6.5, 8.0-8.2)

Dr. J S Chandan, Prof. Medgar Evers College, City University of New York, New York
Units: (2.0-2.2, 2.4, 2.5-2.6, 2.8, 3.6, 9.4, 11.3)

C R Kothari, Former Principal, University Commerce College, Jaipur, and Associate Professor in the Department of Economic Administration and Financial Management (EAFM), University of Rajasthan, Jaipur
Units: (2.3, 3.0-3.2, 9.3)

Dr. Rupesh Tyagi, Assistant Professor(Contractual), Deptt. of Economics, CCS University, Meerut
Unit: (5.4)

Dr. Vidhisha Vyas, Associate Professor and Dean at IILM University Gurugram, Haryana, India
Units: (7, 10, 13, 14)

Vikas Publishing House, Units (2.7, 2.9, 2.10-2.14, 3.3, 3.4, 4.5, 3.7-3.11, 4.3-4.4, 4.7-4.11, 5.0-5.2, 5.5-5.10, 6.0-6.4, 6.6-6.10, 8.3-8.4, 8.5-8.9, 9.0-9.2, 9.5-9.9, 11.0-11.2, 11.4-11.9, 12)



Vikas® is the registered trademark of Vikas® Publishing House Pvt. Ltd.

VIKAS® PUBLISHING HOUSE PVT. LTD.
E-28, Sector-8, Noida - 201301 (UP)
Phone: 0120-4078900 • Fax: 0120-4078999
Regd. Office: A-27, 2nd Floor, Mohan Co-operative Industrial Estate, New Delhi 1100 44
• Website: www.vikaspublishing.com • Email: helpline@vikaspublishing.com

Work Order No. AU/DDE/DE12/04/IV Sem- Printing of Course Material/2021 Dated 21.04.2021 - 400

SYLLABI-BOOK MAPPING TABLE

Econometric Methods

Syllabi	Mapping in Book
BLOCK I: BASIC ECONOMETRICS **Unit-1: SCOPE AND GOALS OF ECONOMETRICS** Definition, Nature and Scope of Econometrics, Goals of Econometrics. **Unit-2: STATISTICAL CONCEPTS** Statistical Concepts, Normal Distribution, Chi-Square, t and F-Distributions, Estimation of Parameters, Properties of Estimators, Testing of Hypothesis.	**Unit-1:** Scope and Goals of Econometrics **(Pages 1-16);** **Unit-2:** Statistical Concepts **(Pages 17-72)**
BLOCK II: LINEAR REGRESSION **Unit-3: SIMPLE LINEAR REGRESSION** Simple Linear Regression, Estimation of Model by Method of Ordinary Least Squares, Properties of Estimators, Goodness of Fit, Tests of Hypothesis, Scaling and Units of Measurement. **Unit-4: MULTIPLE LINER REGRESSION MODEL** Multiple Linear Regression Model, Estimation of Parameters, Properties of OLS Estimators, Goodness of Fit - R2 and Adjusted R2.	**Unit-3:** Simple Linear Regression **(Pages 73-97);** **Unit-4:** Multiple Liner Regression Model **(Pages 98-136)**
BLOCK III: ECONOMETRIC ANALYSIS **Unit-5: VIOLATIONS OF CLASSICAL ASSUMPTIONS** Violations of Classical Assumptions, Consequences, Detection and Remedies Multicollinearity, Heteroscedasticity, Serial Correlation. **Unit-6: SPECIFICATION ANALYSIS** Specification Analysis, Omission of a Relevant Variable, Inclusion of Irrelevant Variable, Tests of Specification Errors. **Unit-7: PANEL DATA MODELS** Panel Data Models, Methods of Estimation, Fixed Effects Model, Random Effects Model. **Unit-8: REGRESSION ON DUMMY VARIABLES** Regression on Dummy Variables, Nature of Dummy Variables, Use of Dummy Variables in Seasonal Analysis and in Combining Time Series, and Cross Sectional Data. **Unit-9: PROBLEMS OF INFERENCE** The Problem of Inference, The Normality Assumption, Hypothesis Testing about Individual Partial Regression Coefficients, Testing the Overall Significance of the Sample Regression. **Unit-10: LINEAR RESTRICTIONS** Linear Restrictions, Testing Joint Hypothesis, Problems and Application using STATA. **Unit-11: ASSUMPTIONS AND SPECIFICATION OF HYPOTHESIS TESTING** Testing of Hypothesis, Assumptions, Specification, Testing of Hypothesis, Prediction, and Applications.	**Unit-5:** Violations of Classical Assumptions **(Pages 137-168);** **Unit-6:** Specification Analysis **(Pages 169-188);** **Unit-7:** Panel Data Models **(Pages 189-202);** **Unit-8:** Regression on Dummy Variables **(Pages 203-243);** **Unit-9:** Problems of Inference **(Pages 244-284);** **Unit-10:** Linear Restrictions **(Pages 285-301);** **Unit-11:** Testing of Hypothesis **(Pages 302-320)**
BLOCK IV: ECONOMETRIC METHODS AND SOFTWARE PACKAGES **Unit-12: ESTIMATION METHOD** Estimation Methods, Single Equation and Systems Estimation Methods, Numerical Problems. **Unit-13: DYNAMIC ECONOMETRIC MODELS** Dynamic Econometric Models, Nature and Preliminary Analysis of Economic Time Series, Integration, Tests of Stationary, Unit Root Test, Non-Stationary and the Problem of Spurious Regression. **Unit-14: ECONOMETRIC SOFTWARE PACKAGE: STATA** Introduction to Econometric Software Package GRETL; E-VIEWS; STATA (any one).	**Unit-12:** Estimation Methods **(Pages 321-340);** **Unit-13:** Dynamic Econometric Models **(Pages 341-359);** **Unit-14:** Econometric Software Package: Stata **(Pages 360-378)**

CONTENTS

INTRODUCTION

NOTES

The 'Econometric Methods' are very useful to estimate the economic variables and to forecast the intended variables, which makes the use of statistical tools and economic theories in combination. Econometrics is the application of statistical methods to economic data in order to give empirical content to economic relationships. More precisely, it is "The quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference". An introductory economics textbook describes econometrics as allowing economists "To sift through mountains of data to extract simple relationships". The first known use of the term "Econometrics" (in cognate form) was by Polish economist Pawe³ Ciompa in 1910. Jan Tinbergen is one of the two founding fathers of econometrics. Ragnar Frisch, also coined the term in the sense in which it is used today.

Applied econometrics uses theoretical econometrics and real-world data for assessing economic theories, developing econometric models, analysing economic history, and forecasting. Econometrics may use standard statistical models to study economic questions, but most often they are with observational data, rather than in controlled experiments. In this, the design of observational studies in econometrics is similar to the design of studies in other observational disciplines, such as astronomy, epidemiology, sociology, and political science. Analysis of data from an observational study is guided by the study protocol, although, exploratory data analysis may be useful for generating new hypotheses.

A basic tool for econometrics is the multiple linear regression model. Econometric theory uses statistical theory and mathematical statistics to evaluate and develop econometric methods. Econometricians try to find estimators that have desirable statistical properties including unbiasedness, efficiency, and consistency. Applied econometrics uses theoretical econometrics and real-world data for assessing economic theories, developing econometric models, analysing economic history, and forecasting.

This book, *Econometric Methods*, is divided into four blocks, which are further subdivided into fourteen units. The topics discussed include definition, nature and scope of econometrics, normal distribution, chi-square, t and F-distributions, estimation of parameters, properties of estimators, testing of hypotheses, simple linear regression, estimation of model by method of ordinary least squares, properties of estimators, goodness of fit, tests of hypotheses, scaling and units of measurement, multiple linear regression model, estimation of parameters, properties of OLS estimators, goodness of fit - R^2 and adjusted R^2, violations of classical assumptions, consequences, detection and remedies multicollinearity, heteroscedasticity, serial correlation, specification analysis, omission of a relevant variable, inclusion of irrelevant variable, tests of specification errors, panel data models, fixed effects model, random effects model, regression on dummy variables,

problem of inference, normality assumption, linear restrictions, testing joint hypothesis, problems and application using STATA, testing of hypothesis, dynamic econometric models, tests of stationary, unit root test, and introduction to econometric software package- STATA.

NOTES

The book follows the Self-Instructional Mode (SIM) wherein each unit begins with an 'Introduction' to the topic. The 'Objectives' are then outlined before going on to the presentation of the detailed content in a simple and structured format. 'Check Your Progress' questions are provided at regular intervals to test the student's understanding of the subject. 'Answers to Check Your Progress Questions', a 'Summary', a list of 'Key Words', and a set of 'Self-Assessment Questions and Exercises' are provided at the end of each unit for effective recapitulation.

BLOCK - I
BASIC ECONOMETRICS

NOTES

UNIT 1 SCOPE AND GOALS OF ECONOMETRICS

Structure

1.0 INTRODUCTION

It is essential to keep in mind that the application of econometrics is not just restricted to the economics domain. In fact, it has very widespread application. It is used in pure science domains as also social sciences. It may be pointed out that the econometric methods can be used whenever there is a need of finding a stochastic relationship in mathematical format. Econometric tools aid in expounding relationships among variables.

Econometrics makes use of data derived in various ways. Some of these are cross sectional data (at a point in time, various economic units are observed), time series data (one or more variables are observed over a period of time), pooled cross sections (data set comprising time series and cross-sectional features) and panel or longitudinal data (comprises a time series for every one of the cross-sectional member in the data set).

The econometric methodology comprises eight steps to be followed in the specified order, starting with the statement of theory or hypothesis and successfully culminating with the application of the model for control or policy purpose.

When the four branches of economics merge together, it becomes econometrics. Both a science and an art, the goal of econometrics is forecasting. Econometrics is considered to be an art mainly because the data it uses is generally incomplete and unobserved for validating a hypothesis and due to this human creativity is required to strike a balance between realistic approximation and scientific rigor.

According to Hansen (1996): "*Econometrics is alchemy since econometricians can create nearly any result desired, but it is also science because econometricians also know how to reject and avoid spurious models*".

In this unit, you will study about the definition of econometrics, nature and scope of econometrics, and goals of econometrics.

1.1 OBJECTIVES

After going through this unit, you will be able to:

- Define the econometrics
- Understand the nature and scope of econometrics
- Analyse the goals of econometrics

1.2 ECONOMETRICS

Econometrics is a branch of economics in which measurement of relationships is discussed. Econometric is the application of a statistical method to economic data in order to give empirical content to economic relationships.

Economic theories attempt to define the quantitative relationships between the different economic variables. These relationships are, at times, described in mathematical terms which help to better understand the economic world we live in. Theories have to be checked against data collected from the real world. If empirical data verify the relationship proposed by the theory, we may accept it; otherwise we must reject the relationship.

In other words, if the theory is compatible with the actual data, we accept the theory as valid. If the theory is not compatible with the actual date we either reject the theory or modify it. Thus, econometric is the analysis and testing of economic theories to verify hypothesis and improve prediction for future.

1.2.1 Meaning of Econometrics

The term econometric was introduced in 1926 by Ragnar Frisch, a Norwegian economist and statistician. In fact, the term was modelled on the expression of Biometric which appeared late in the nineteenth century to denote the field of

biological studies employing statistical methods. Econometric is a set of mathematical and statistical tools that allow describing or testing of the different economic theories and concepts.

Econometrics is a combination of economics, mathematics and statistics. It provides numerical values for parameters of economic relationships and verifying economic theories. The word econometric is made of two components Econo + metric. Econo stand for economics and metric stands for measurement. Thus, the word econometric indicates measurement of economics or economic measurement. In econometrics, mathematical and statistical tool are used to measure the validity of economic theory. Econometric is the science. It is the combination of economic theory with economic statistics in which mathematical and statistical methods are used to investigate the empirical support of the general schematic law established by economic theory. For instance, law of demand reveals that there is inverse relationship between price and demand. When price increases, demand for the commodity decreases, when price decreases demand increases. We verify this theory by a regression, a mathematical and statistical tool to analyse the theory.

1.2.2 Definition of Econometrics

Although measurement plays an important role in econometrics, the scope of econometrics is much broader as described by different econometricians. Different leading econometricians define it in their different manner. Some definitions are given below.

According to Arthur S. Goldberger: "*Econometrics may be defined as the social science in which the tools of economic theory, mathematics and statistical inference are applied to analyse the economic phenomena*."

According to Samuelson, Koopmans, and Stone: "*Econometrics may be defined as the quantitative analysis of actual economic phenomena based on current development of theory and observation, related to appropriate method of inferences*."

According to A. S. Goldberger: "*The main task of econometric theory is to provide a bridge between the exact relationships of economic theory and the disturbed relationships of economic reality*."

According to H. Theil: "*Econometric is concerned with the empirical determination of economic laws*."

In simple words, econometrics may be considered as the integration of economics, mathematics and statistics for the purpose of providing numerical values for parameters of economic relationships and verifying economic theories. In this way, it is a special type of economic analysis and research in which general economic theory formulated in mathematical terms is combined with empirical measurement of economic phenomena.

NOTES

1.3 METHODOLOGY OF ECONOMETRICS

Which method is applied by econometrician to proceeds an economic problem in their analysis. Econometricians solve a specific economic problem by using a specific process. First, an economic theory is chosen and converted into a mathematical model. This mathematical model is converted into an econometric model. Then, the variables are identified by using cross sections and time series observations.

Using these the econometric model is estimated. After the model is estimated, it is put to test at the first level of statistical test. After the first set of testing of validation, the model is subject to the second level of econometric test. If model tests as being valid in the second level of econometric testing, then the model is adequate to proceed to the next level. If the model is not adequate, then it is sent back for reformulation and has to go through all of the steps, as aforementioned in the methodology of framing an econometric model and the proceeding steps. If the theoretical hypothesis testing is accepted, then accept the theory otherwise reject the theory and move to either reformulate the theory or move to alternative theory to verify the theory.

The econometric methodology comprises eight steps to be followed in the specified order. These are:

1. Statement of theory or hypothesis
2. Specification of mathematical model of the theory
3. Specification of the statistical or econometric model
4. Obtaining the data
5. Estimation of the parameters of the econometric model
6. Hypothesis testing
7. Prediction or forecasting
8. Application of the model for control or policy purpose

Following is an explanation of the eight steps:

First Step - Statement of Theory or Hypothesis: Economics gives qualitative statements. Here, we have taken the theory of consumption function. For example, Keynes stated that *the fundamental psychological* law is that as income increases consumption also increases but not at same rate, this is the theory of consumption function. To explain further, Keynes said that the *marginal propensity to consume* (MPC), the rate of change in consumption for a unit change in income, is greater than 0 but less than 1.

Second Step - Specification of Mathematical Model of the Theory: Model formation is the second step. In the second step, the theory is converted into a mathematical model. Here, the econometrician selects variables, and specifies relationships based on economic theories or hypothesis. Next, this theory of

NOTES

consumption function is converted into a mathematical model. Now, there are two variables: one is income and one is consumption.

$$C = \beta_1 + \beta_2 Y;\ 0 < \beta_2 < 1$$

Here, C = Consumption expenditure

Y = Income

β_1 = intercept or autonomous consumption

β_2 = slope coefficient or Induce consumption

Here, β_2 represents MPC; it will show that when income increases how much change will be there in consumption expenditure.

Third Step - Specification of the Statistical or Econometric Model: In this step, the mathematical model is converted into an econometric model. Mathematical model assumes that there is an exact relationship between two variables; it means that consumption expenditure depends only on income. But the relationship between two variables is, generally, not exact. As we know, consumption is not only affected by income it is also affected by some other factors, such as, trend, taste, brand and time. In an econometric model, these other factors are referred to as disturbance terms, error terms or random terms. In other words, we can say that in a mathematical model some relevant explanatory variables are not included; irrelevant explanatory variables are included and some are under-identification. These are denoted by U. The econometrician modifies this inexact consumption function as below:

$$C = \beta_1 + \beta_2 Y + U;\ 0 < \beta_2 < 1$$

Here, U is an error term or random term (stochastic) variable.

Fourth Step - Obtaining the Data: In this step, data of income and consumption are gathered according to sample.

Year	Income (Y) $ billion	Consumption expenditure (C) $ billion
2016	200	160
2017	300	200
2018	400	320
2019	500	360

Fifth Step - Estimation of the Parameters of the Econometric Model: In this step, after collecting the data, the econometrician estimates the parameters of the econometric model. For the purpose of estimation, regression tools are used to obtain the estimation. By using this technique, suppose we obtain:

$$C = \beta_1 + \beta_2 Y$$

$$C = 8 + 0.72Y$$

NOTES

As per this estimation, $\beta_1 = 8 \ and \ \beta_2 = 0.72$, We know that represents to MPC, which is about 0.72. This means that for the sample period when real income increases then real consumption expenditure increased by about 72 percent and the balance 28 per cent is used for saving or investment.

Sixth Step - Hypothesis Testing: An econometrician does not trust these results blindly. The econometrician tests the hypothesis through six different 6 steps to know if MPC<1 is statistically significant or true. If so, it may support the Keynes theory.

In this step, the econometrician evaluates the theory or hypothesis; rejects the economic theory if the theory is not validated, or accepts the economic theory if the theory is validated.

Seventh Step - Prediction or Forecasting: If the hypothesis is true, then the result is used for forecasting or prediction. For example, suppose income = $ 800 billion in 2021 as forecasted consumption expenditure:

$C = \beta_1 + \beta_2 Y$

C= 8+0.72 (600)

C= 584

So, the consumption will be $584 billion.

Eighth Step - Application of the Model for Control or Policy Purpose: These results or forecasts are used for policy purpose by the government. A government uses these when formulating fiscal policy, monetary policy, etc. Suppose that a government believes that consumption expenditure of about $700 billion will keep the unemployment rate at its current level. What level of income will guarantee the target amount of consumption expenditure, if the government wants to continue the same unemployment rate? So, simple arithmetic will show:

$C = \beta_1 + \beta_2 Y$

700= 8+0.72 Y

Y= 961.11

1.3.1 Why is Econometric a Separate Discipline?

Econometrics is an estimation of economic law. There are four branches of economics:

- Economic theory
- Mathematical economics
- Economics statistics
- Mathematical statistics (tools)

If these four branches integrated into one branch it becomes econometrics. It is a separate subject or new product. Econometric is a science of estimation of economic law.

In **Economic Theory**, hypothesis or statements are mostly qualitative in nature. For instance, microeconomic theory states that other things remaining same, decrease in the price of a commodity increases the quantity demand for that commodity. Thus, economic theory suggests a negative or inverse relationship between price and quantity demanded of a commodity. But the theory does not provide any numerical measurement or the relationship between the two; it does not tell by how much demanded quantity goes up or goes down as a result of a certain change in the price of the commodity. It is the job of the econometrician to provide such numerical estimates. Econometric gives empirical content to most economic theory.

In **Mathematical Economics**, economic theory is expressed in mathematical form or equations. Econometrics is mainly concerned with the empirical verification of economic theory. The econometrician often uses the mathematical equations proposed by the mathematical economist and puts these equations in such a form that they give themselves to empirical testing.

A great deal of ingenuity and practical skill is required for the conversion of mathematical equations to econometric equations.

Economic Statistics is mainly concerned with collecting, processing and presenting economic data in the form of tables and charts. Primarily, an economic statistician is responsible for collecting data from the population. These collected data are the raw material (data) for econometric work. An economic statistician is not concerned with using collected data to test hypothesis or economic theory. Of course, the person who does that becomes an econometrician. An econometrician uses this data to test the validity of hypothesis of economic theory.

Mathematical Statistics provides tools for estimation. These tools are applicable on different condition. For example, when using these tools in biological science, it is referred to as biometric. In biology, data are generated in a controlled manner, but in social science, especially in economics, we cannot generate data in a controlled manner. All data are changing simultaneously, so using mathematical statistics in an uncontrolled environment becomes econometric. An econometrician often needs special methods in view of the unique nature of most economic data, due to the fact that data are not generated in a controlled condition.

1.3.2 Econometrics and Mathematical Economics

In mathematical economics, the literary forms of economics are used in terms of mathematical symbols. Mathematical economics is only an approach to economic analysis and does not differ from the non- mathematical approach to economic analysis in any fundamental way. The major difference between the mathematical economics and literary economics lies principally in the fact that in mathematical economics assumptions and conclusion are stated in mathematical symbols and equations, while in literary economics these are stated in words, sentences or statements. Both describe the same relationships.

Neither economic theory nor mathematical economics allows for random elements which might affect the relationships and make it stochastic in character. Relations in mathematical economics or in economic theory are of non-stochastic form. Although econometrics presumes the economic theory in the mathematical form, it does not assume that economic relations are exact (non-stochastic).

Econometric assumes every economic relation as a stochastic. There are many causes for assuming the presence of disturbance term in the exact relations to make it stochastic.

Econometric methods provide numerical value for the coefficients of the relationships under investigation. While mathematical economics does not provide such numerical value, econometric enables us to pass from the abstract theoretical scheme to numerical result by combining mathematical formulations of theory with empirical data in concrete cases. Thus, econometric provides a bridge between the exact relations of economic theory and disturbed relations of economic reality.

1.3.3 Econometrics and Statistics

Statistics deals with the collection of data and its tabulation in a desired form. Economic statistics is mainly a descriptive aspect of economic theory. In case of mathematical economics, it does not provide numerical values of the parameters involve in the economic relationships.

Economic statistics also differ from mathematical statistics. The fundamental mathematical statistics are applicable in econometric, but they are not applied blindly. They are used only after adapting them to the random or stochastic behaviour occurring in economic problems. These adapted statistical methods are then called econometric methods.

1.4 GOALS OF ECONOMETRICS

Mathematical economics and economic statistics are the important aspects of econometrics. Mathematical formulation of theory provides determination and accuracy, while statistics provides the life blood or raw date to the new hybrid field knowledge of econometrics. The prime goals of econometrics are:

1. Verification of economic theory or judging the validity of the economic theory,
2. Estimation of coefficient of economic relations,
3. Forecast the future value of economic magnitude.

Verification of Economic Theory: Primary goal of econometrics is the verification of economic theories. No theory can stand on its own merits without some empirical testing and econometrics allows us to make such empirical testing. Econometric helps us to know and decide how well economic theories expound and so far explain the actual behaviour of the economic units.

Estimation of Coefficients of Economic Relations: Econometric is concerned with the analysis of measures of economic activities. Various economic techniques are applied in order to obtained reliable estimates of the individual coefficients of the economic relationships with which different parameters of economic theory may be evaluated. For example, econometrics can provide estimates of marginal costs, marginal revenue, elasticity of demand, and elasticity of supply, multiplier coefficients, and technical coefficients of production, etc. The knowledge of all such coefficients is extremely valuable for the formulation of sound economic policies.

Estimation of production function enables us to compare marginal productivity of labour and capital in various firms and industries; and also provides an insight of returns to scale the industry is experiencing.

Forecasting the Future Value of Economic Magnitudes: After the evaluation of the economic theory and estimation of numerical values of the coefficients of economic relations, the econometric models can be used for forecasting. At present, forecasting is gradually becoming more important both for the regulation of developed economies as well as for the planning for the economic development of underdeveloped or undeveloped economies.

Forecasting is the estimation or observation of the expected value of the dependent variable for the given value of independent variable. For example, from the estimate of the marginal propensity to consume, the investment multiplier can be computed by a simple formula - $K = \frac{1}{1-MPC}$.

Let MPC = 0.7

K=1/(1-MPC)   K=1/(1-0.7)   K = 1/0.3   K = 3.33

Then, for a given increase in investment, the net ultimate increase in national income. If the forecast value of increase in national income turns out to be lower than what is desired, the government must take different measures in order to achieve the target. Thus, the forecast enables the government, or policy makers, to judge if it is necessary to take any measure in order to influence the relevant economic variable.

1.5 NATURE OF THE ECONOMIC APPROACH

Nature of econometrics refers to whether it is a science or an art. Econometric is a science, because like in other science discipline in econometrics, too, all procedures are carried out in a systematic (scientific) manner. In a scientific study, we first observe several things or behaviours and only then data are gathered followed by the analysis of the data, and finally conclusions are arrived at.

Similarly, in econometric analysis also, we first observe certain behaviours, particularly some social behaviour. Then, we collect data pertaining to these

behaviours. After collecting the data, we analyse these data sets and come to certain conclusion. The main objective of data analysis is to come up with a conclusion so that we can conclude whether the particular theory is valid in a particular context or not valid.

Econometric can also be an art because like any other artist an econometrician analyses data. In econometric we learn how to set economic models, estimate variables and test the validity of an economic theory.

It can also be said that, basically, econometric is the science and art of using economic data, the economic data generated from economic decision-making. Economic data can be used in several different contexts. Even this particular technique can be used to analyse and provide empirical validity to economic theory.

Econometric is both science and art. There are some fixed stages which every researcher must take into account:

Stage 1 - first and foremost step of every econometric research is to take up any statement of theory or hypothesis.

Stage 2 - this stage of research is the specification of the model. The specifications of the model have to be based on economic theory. In this stage the following have to be finalised:

i. The dependent and independent variables to be included in the model
ii. The priori assumptions about the size and the sign of the parameter of the model
iii. The mathematical formation of the model

Stage 3 – in this stage estimation of the model by means of appropriate econometric method is to be done. This include:

i. Collection of data on the variables included in the model
ii. Examination of the multi-collinearity problem
iii. Examination of the identification conditions if the model involves more than a single equation
iv. Choice of appropriate econometric technique for estimation of the model

Stage 4 – in this stage, the estimated model is to be evaluated on the basis of certain criteria to ascertain whether the estimates are reliable.

The evaluation consists of deciding whether the estimates of the parameters are theoretically meaningful and statistically significant. The following three criteria are used for such evaluation.

Economic 'A Priori' Criteria: These are determined by the principles of economic theory. If the estimates of the parameters turn up with sign and size not conforming to a priori criteria, they should be rejected unless there is a good reason to believe that in the particular case the principles of economic theory do not hold.

In such cases, the reasons for accepting the estimates with the different sign and size must be stated clearly.

Statistical Criteria (first-order test): These tests are determined by statistical theory. They include standard deviation or standard error () and correlation coefficient () of the estimates. Coefficient of determination (r^2) computed from the sample data explains the percentage of the total variations in the dependent variable due to change in explanatory variables. The standard deviation () of the estimates describes the dispersion of the estimates around the true parameter. Hence, the larger the standard error, the less reliable it is, and vice versa.

Econometric Criteria (second-order test): Theory of econometric gives these tests. These tests help to establish whether the estimates have the desirable properties of unbiasedness, sufficiency and consistency. If the assumptions of the econometric technique applied are not satisfied, the estimates cease to possess some of the desirable properties. The assumptions of the various econometric techniques differ and, hence, there are various econometric criteria for each method. Before accepting or rejecting the estimates, it is essential to use all the above criteria.

The major objective of econometrics is forecasting. In the final stage, the forecasting power of the model is to be examined. Forecasting is closely related to policy choice and policy evaluation. In fact most methods of policy evaluation rely upon a specific type of forecast. Many times, the model may be economically meaningfully and statistically and economically correct for the sample period for which the model has been estimated, yet it may possess very bad forecasting power. This may be due to sensitiveness of the structural parameters involved in the model or due to the value of explanatory variable not being accurate; or the estimates of the coefficients not being correct.

The estimated value (forecast value) is than compared with the actual magnitude of the relevant dependent variable. The difference between two variables is tested statistically. If after conducting the desired test of significance it is observed that this difference is significant, it is to be concluded that the forecasting power of the model is poor.

Econometrics is classified into two branches:

- Theoretical
- Applied

Theoretical econometrics deals with the development of the appropriated methods for measuring economic relationships described by an econometric model. Theoretical econometrics also spells out the assumption of these methods and their properties. It is concerned with what happens to these properties when one or more of the assumptions of the method are not fulfilled.

Applied econometrics describes the practical value of econometric research. It deals with the applications of econometric techniques developed in theoretical econometrics to different fields of economic theory for its verification and

NOTES

forecasting. Presently, in the fields of market demand and supply, cost function, and consumption and investment functions econometrics is being vastly employed. Applied econometrics has made it possible to obtain numerical results from these studies which are very important for policy makers.

Check Your Progress

1. What does the terms 'Econo' and 'Metric' stand for?
2. Tools derived from which disciplines are used in econometrics to measure the validity of economic theory?
3. Who defined econometrics in the following terms?

   "Econometrics may be defined as the social science in which the tools of economic theory, mathematics and statistical inference are applied to analyse the economic phenomena."
4. According to A. S. Goldberger, what is the main task of econometric theory?
5. What can be considered the purpose of econometrics being an integration of economics, mathematics and statistics?
6. Explain the first step in the econometric methodology.
7. At what stage can econometric models be used for forecasting?
8. In the econometric methodology, which step follows after the specification of mathematical model of the theory?
9. With what is economic statistics mainly concerned?
10. Illustrate the two branches of econometrics.

1.6 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The term 'Econo' stand for economics and 'Metric' stands for measurement.
2. In econometrics, mathematical and statistical tool are used to measure the validity of economic theory.
3. Arthur S. Goldberger.
4. According to A. S. Goldberger, the main task of econometric theory is to provide a bridge between the exact relationships of economic theory and the disturbed relationships of economic reality.
5. Econometrics may be considered as the integration of economics, mathematics and statistics for the purpose of providing numerical values for parameters of economic relationships and verifying economic theories.

6. The first step in the econometric methodology is the creation of the statement of theory or hypothesis.
7. After the evaluation of the economic theory and estimation of numerical values of the coefficients of economic relations, the econometric models can be used for forecasting.
8. In the econometric methodology, specification of the statistical or econometric model is conducted after the specification of mathematical model of the theory.
9. Economic Statistics is mainly concerned with collecting, processing and presenting economic data in the form of tables and charts.
10. The two branches of econometrics are theoretical and applied.

1.7 SUMMARY

- Econometrics is a branch of economics in which measurement of relationships is discussed. It is the application of statistical method to economic data in order to give empirical content to economic relationships.
- Economic theories attempt to define the quantitative relationships between the different economic variables.
- If theory is compatible with the actual data, we accept the theory as valid. If the theory is not compatible with the actual date we either reject the theory or modify it.
- Econometric is a set of mathematical and statistical tools that allow describing or testing of the different economic theories and concepts. It is a combination of economics, mathematics and statistics.
- The word econometric is made of two components Econo + metric. Econo stand for economics and metric stands for measurement. Thus, the word econometric indicates measurement of economics or economic measurement.
- Econometrics used mathematical and statistical tool to measure the validity of economic theory. Econometric is both a science and an art.
- In simple words, econometrics may be considered as the integration of economics, mathematics and statistics for the purpose of providing numerical values for parameters of economic relationships and verifying economic theories.
- In the econometric methodology, there are eight specific stages, which are: statement of theory or hypothesis, specification of mathematical model of the theory, specification of the statistical or econometric model, obtaining of data, estimation of the parameters of the econometric model, hypothesis testing, prediction or forecasting, and application of the model for control or policy purpose.

NOTES

- The major objective of econometrics is forecasting. Forecasting is closely related to policy choice and policy evaluation. Many times, a model may be economically meaningfully and statistically and economically correct, yet it may possess very bad forecasting power. This may be due to sensitiveness of the structural parameters involved in the model or due to the value of explanatory variable not being accurate; or the estimates of the coefficients not being correct.
- The major difference between the mathematical economics and literary economics lies principally in the fact that in mathematical economics assumptions and conclusion are stated in mathematical symbols and equations, while in literary economics these are stated in words, sentences or statements.
- Statistics deals with the collection of data and its tabulation in a desired form. Fundamental mathematical statistics are applicable in econometric, but they are not applied blindly. They are used only after adapting them to the random or stochastic behaviour occurring in economic problems. These adapted statistical methods are then called econometric methods.
- The prime goals of econometrics are: verification of economic theory or judging the validity of the economic theory, estimation of coefficient of economic relations, and forecast the future value of economic magnitude.
- Applied econometrics describes the practical value of econometric research. It deals with the applications of econometric techniques developed in theoretical econometrics to different fields of economic theory for its verification and forecasting.

1.8 KEY WORDS

- **Econometrics:** Econometrics is a branch of economics in which measurement of relationships is discussed. It is the application of statistical method to economic data in order to give empirical content to economic relationships.
- **Empirical content**: content which is derived from or relating to experiment and observation rather than theory.
- **Quantitative relationship**: the relation between things (or parts of things) with respect to their comparative quantity, magnitude, or degree.
- **Methodology**: a way of doing something based on particular principles and methods.
- **Forecasting**: to define with the help of specific information what will probably happen in the future.

- **Statistics:** Statistics deals with the collection of data and its tabulation in a desired form. Fundamental mathematical statistics are applicable in econometric, but they are not applied blindly.

1.9 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What is the statement of theory or hypothesis?
2. According to A. S. Goldberger, what is the main task of econometric theory?
3. Explain the prediction or forecasting.
4. Define the four branches of economics.
5. State the methodology used by econometrics.

Long-Answer Questions

1. Explain the nature of econometrics. How it is both a science and an art?
2. Define the eight steps of the econometric methodology with the help of examples.
3. Write in your own words what is econometrics?
4. Analyse the relationship between Econometric and Mathematical Economics.
5. Describe the relevance of economic theory as a subject. Give appropriate examples.
6. Elaborate on the goals of econometrics.

1.10 FURTHER READINGS

Johnston, J. and John DiNARDO. 1997. *Econometric Methods*, Fourth Edition. New Delhi: Tata McGraw-Hill.

Koutsoyiannis, A.1977. *Theory of Econometrics,* Second Edition. London: The Macmillan Press Ltd.

Özdemir, Durmuº. 2016. *Applied Statistics for Economics and Business,* Second Edition. Izmir (Turkey): Springer.

Maddala, G. S. 1992. *Introduction to Econometrics,* Second Edition. New York: Macmillan Publishing Company.

NOTES

Pindyck, R. S and D. L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts,* Fourth Edition. New York: McGraw Hill.

Goldberger, A. S. 1998. *Introductory Econometrics***.** Cambridge: Harvard University Press.

Levine, David M., Timothy C. Krehbiei, Mark L. Berenson and P. K. Viswanathan. 2009. *Business Statistics*, Fifth Edition. New Delhi: Pearson Education.

Webster, Allen L. 1998. *Applied Statistics for Business and Economics,* Third Edition. New Delhi: Tata McGraw-Hill.

UNIT 2 STATISTICAL CONCEPTS

Structure

2.0 INTRODUCTION

Econometrics is the application of statistical methods to economic data in order to give empirical content to economic relationships. More precisely, it is 'The quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference'.

Normal distributions are important in statistics and are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known. Their importance is partly due to the central limit theorem.

A Chi-squared test, also written as $\chi 2$ test, is a statistical hypothesis test that is valid to perform when the test statistic is Chi-squared distributed under the null hypothesis, specifically Pearson's Chi-squared test and variants thereof.

In probability and statistics, Student's *t*-distribution (or simply the *t*-distribution) is any member of a family of continuous probability distributions that arise when estimating the mean of a normally-distributed population in situations where the sample size is small and the population's standard deviation is unknown. In probability theory and statistics, the *F*-distribution, also known as Snedecor's

F-distribution or the Fisher–Snedecor distribution is a continuous probability distribution that arises frequently as the null distribution of a test statistic, most notably in the ANalysis Of VAriance (ANOVA) and other F-tests.

NOTES

Parameters are descriptive measures of an entire population that may be used as the inputs for a Probability Distribution Function (PDF) to generate distribution curves. Estimation theory is concerned with the properties of estimators; that is, with defining properties that can be used to compare different estimators (different rules for creating estimates) for the same quantity, based on the same data. Such properties can be used to determine the best rules to use under given circumstances.

The goodness of fit of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. Such measures can be used in statistical hypothesis testing. A statistical hypothesis is a hypothesis that is testable on the basis of observed data modelled as the realised values taken by a collection of random variables. A set of data is modelled as being realised values of a collection of random variables having a joint probability distribution in some set of possible joint distributions. The hypothesis being tested is exactly that set of possible probability distributions. Measurement scale, in statistical analysis, the type of information provided by numbers. Each of the four scales (i.e., nominal, ordinal, interval, and ratio) provides a different type of information. Measurement refers to the assignment of numbers in a meaningful way, and understanding measurement scales is important to interpreting the numbers assigned to people, objects, and events.

In this unit, you will study about the statistical concepts, normal distribution, Chi-square, t and F- distribution, estimation of parameters, properties of estimators, testing of hypotheses.

2.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the statistical basic concepts
- Comprehend the normal distribution
- Explain about the Chi-square
- Analyse the t and F- distribution
- Discuss about the estimation of parameters
- Elaborate on the properties of estimators
- Interpret the testing of hypotheses

2.2 BASIC CONCEPT OF STATISTICAL

NOTES

Econometrics is the application of statistical methods to economic data in order to give empirical content to economic relationships. More precisely, it is 'The quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference'. An introductory economics textbook describes econometrics as allowing economists 'To sift through mountains of data to extract simple relationships'. The first known use of the term 'Econometrics' (in cognate form) was by Polish economist Pawe Ciompa in 1910.

A basic tool for econometrics is the Multiple Linear Regression (MLR) model. Econometric theory uses statistical theory and mathematical statistics to evaluate and develop econometric methods. Econometricians try to find estimators that have desirable statistical properties including unbiasedness, efficiency, and consistency. Applied econometrics uses theoretical econometrics and real-world data for assessing economic theories, developing econometric models, analysing economic history, and forecasting.

Theory

Econometric theory uses statistical theory and mathematical statistics to evaluate and develop econometric methods. Econometricians try to find estimators that have desirable statistical properties including unbiasedness, efficiency, and consistency. An estimator is unbiased if its expected value is the true value of the parameter; it is consistent if it converges to the true value as the sample size gets larger, and it is efficient if the estimator has lower standard error than other unbiased estimators for a given sample size. Ordinary Least Squares (OLS) is often used for estimation since it provides the 'Best Linear Unbiased Estimator' (BLUE) (where 'Best' means most efficient, unbiased estimator) given the Gauss-Markov assumptions. When these assumptions are violated or other statistical properties are desired, other estimation techniques, such as maximum likelihood estimation, generalised method of moments, or generalised least squares are used. Estimators that incorporate prior beliefs are advocated by those who favour Bayesian statistics over traditional, classical or 'Frequentist' approaches.

2.2.1 Descriptive Statistics

As the name suggests, descriptive statistics merely describes the data and consists of methods and techniques used in collection, organization, presentation and analysis of data in order to describe the various features and characteristics of such data. These methods can either be graphical or computational. Thus, data can be presented in the form of a chart or a table in order to show certain trends, proportions, maximum and minimum values, and so on. For example, if we simply describe the number of workers in different types of industries in America, then

NOTES

that would constitute descriptive statistics. In addition to the organization of data, the field of descriptive statistics is concerned with the analysis of data so that the data can be easily understood. Averages, proportions and other measures that describe the spread of data around the average are also some of the measures used to describe data. By using these measures, we summarize the data and even though we may lose the detail, we gain clarity and compactness. For example, the following statistics, in their most summarized presentation describe in some way the characteristics of the population from which they were drawn:

- The ages of students in my statistics class range from 19 – 45 years.
- The average IQ of students at our college is 140.
- 20 per cent of the students in my class are married.

All these examples simply summarize and describe the data. Not much can be inferred from them, nor can definite decisions be made or conclusions drawn.

For a proper appreciation of the various descriptive statistics involved, it is necessary to note that most of the statistical distribution have some common features. Though the size of the variables varies from item to item, most of the items are distributed in such a manner that if we move from the lowest value to the highest value of the variable, the number of items at each successive stage increases with a certain amount of regularity till we reach a maximum; and then as we proceed further, they decrease with the similar regularity. If we plot the percentage frequency density, i.e., the percentage of cases in an interval of unit variable width, we get frequency curves of the type shown in Figure 2.1. (Note that the area under each curve should be equal to 100, the total percentage points).

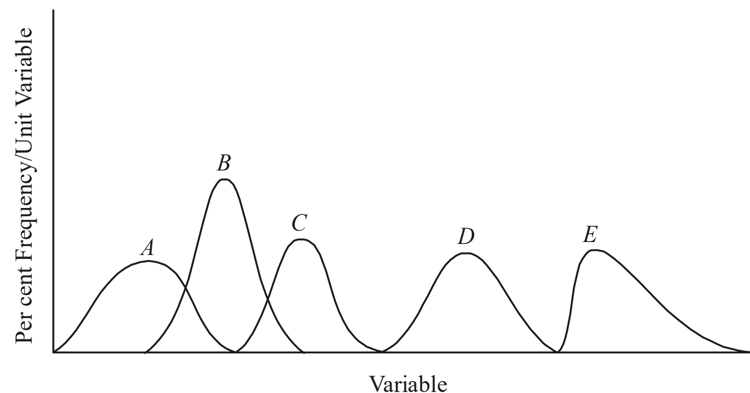


Fig. 2.1 *Representation of Measures of Central Tendency*

There are various 'gross' ways in which frequency curves can differ from one another. Even when the 'general' shapes of the curves are the same (the area under them already made equal by the strategy of plotting the per cent density), the details of the shape may change. Thus, curve *B* has a smaller spread than *A*, curve *C* is more peaky and curve *E* is less symmetrical. Even when the

curves have almost the same shape (i.e., same spread, peakiness, symmetry, etc.) as in curves *A* and *D*, the two may differ in location along the variable axis. Thus, the items of distribution *D* are generally larger than those of *A*; so are those of *B* compared to *A*. Thus, a kind of an 'average' location of the distribution along the variable axis is an important descriptive statistics. These statistics are collectively known as measures of location or of central tendency.

2.2.2 Inferential Statistics

Inferential statistics can be defined as those methods that are used to estimate a characteristic of a population or making a decision concerning a population on the basis of results obtained from a sample taken from the same population. The measured characteristics of the sample are known as sample statistics, while the measured characteristics of the population are known as population parameters. A major portion of statistics deals with making decisions, inferences, predictions and forecasts about the population based on the results obtained from samples taken from such populations.

The need for inferential statistical methods derives from the need for sampling. As the population becomes large, it is usually too costly, too time consuming and too cumbersome to take the entire population into consideration in order to obtain any information of interest. Of course, the results obtained from the entire population are the most accurate and if the population is indeed small, then it is advisable to consider the entire population. However, when the population is large—sometimes considered infinite—then sampling method is used.

The question is: How do these sample statistics relate to population parameters? Can we state that the conclusions drawn from the analysis of the sample are exactly the same as the conclusions that would be drawn from the entire population from which the representative sample was taken? The answer is unlikely. How close is the sample characteristics to the population characteristics would depend upon the randomness of the sample as well as the size of the sample. The more random the sample is and larger the sample is, the more closely its characteristics would be with the population characteristics. This link, in terms of the degree of closeness, is provided by probability theory. Probability theory provides the link by ascertaining the likelihood that the results from the sample reflect the results from the population.

Our interest is not in finding the characteristics of a sample, but to find the characteristics of the population. Sampling is simply a means to the end. For example, when we say that we want to know the salary of university professors, we mean the salary of all university professors and not simply of the sample we have taken. Only then can observations and decisions be made in this regard. Similarly, if we want to know what percentage of eligible voters will vote for a specific political party in the next general elections in India, a sample in itself would not indicate that, and we cannot ask the entire population. Our decisions and

NOTES

projections would be based on the inclination of the entire population. A sample in itself would not mean much. However, if the sample truly represents the population, then we can draw conclusions about the population on the basis of sample results. Appended to these conclusions will be a probability statement specifying the likelihood or confidence that the results from the sample reflect the voting behaviour of the population. Usually, the margin of error is stated as ± 3 – 5 per cent.

The following are some of the situations that the field of inferential statistics deals with:

(i) Between 35 per cent and 40 per cent of graduate students in the universities are married. These statistics refer to the entire population of graduate students. It would be reasonable to assume that these percentages were calculated on the basis of samples taken from the population of all graduate students. The students in these samples were asked in order to know as to how many of these students were married. The answers formed the basis for drawing conclusions about the entire population of the graduate students.

(ii) There is a definitive association between smoking and lung cancer. This statement is the result of endless research on many samples taken and studied in order to find out if there is any correlation between smoking and lung cancer, and based upon the results thus obtained from sample studies, a valid statement about the association of smoking with lung cancer in the whole population can be made.

(iii) 30 per cent of all television viewers watched the show '20/20' last night. This statement can be compared with the following statement: 30 per cent of those who were interviewed watched the show '20/20' last night. The latter statement is descriptive statistics since it only presents the data in a summarized form. However, if we infer from the second statement to reach at the first statement, then the first statement is an example of statistical inference.

(iv) Suppose that the Vice-chancellor of MG University wanted to conduct a survey to learn about student perceptions concerning the quality of life on campus. The population will be all the students enrolled in the university, while a sample will consist of only the students who have been randomly selected to be included in the sample to participate in the survey. The goal is to determine the various attitudes and characteristics of interest relating to quality of student life in the entire university by using the sample statistics to draw conclusions about the similar population characteristics

(v) Between 35 per cent and 40 per cent of graduate students in the universities are married. This statistics refers to the entire population of graduate students. It would be reasonable to presume that these percentages were calculated on the basis of samples taken from the

population of all graduate students. The students in these samples were asked in order to know how many of these students were married. The answers formed the basis for drawing conclusions about the entire population of graduate students.

NOTES

2.2.3 Statistical Enquiry

Statistical enquiry refers to an investigation of a given phenomenon on the basis of statistical and quantitative models and techniques. An enquiry is defined as a close examination of a matter in a search of information or truth. 'Close' here means intense and comprehensive. The matter must be looked in depth and not merely at the surface. Close examination requires focus as well as dedication. The matter must not only be examined, but also analysed. Sufficient and adequate information about the matter under study must be made available for meaningful conclusion. Search for truth simply means search for facts as they relate to the matter under consideration. Facts must be consistent with our perception of reality. Truth necessarily requires absence of bias or prejudice in the process of investigation.

Statistical enquiry involves objective analysis of information in order to arrive at some meaningful conclusion. For example, the Federal Drug Administration (FDA) approves a drug for consumption after lengthy statistical investigations regarding the results of the experiments about the need and usefulness of the drug. Similarly, statements regarding inflation rates or unemployment rates are arrived at after significant statistical data and analysis of such data. These conclusions must be as accurate as possible for them to become the foundations for national economic policies.

Non-statistical enquiry, on the other hand, is gathering of opinions and feelings, and the conclusions so arrived at are more subjective in nature. It is more of an observation rather than an enquiry. Observations and statements about beauty, goodness, honesty, and so on are all non-statistical enquiries. A person's views and opinions about an issue such as abortion or whether the Congress party governs the country better than the BJP or vice versa are also non-statistical enquiries. Even though at an individual level such opinions remain in the domain of non-statistical enquiry, collectively they may be subjected to statistical enquiry. For example, if significantly large data is collected regarding the opinions of people about governance by the Congress or the BJP over the years, a statistical pattern can be developed and statistical analysis can be made for any changes in the opinions over the given time period. This, then , would constitute statistical enquiry.

Statistical enquiry is more reliable as it deals with quantitative data and quantitative tools and methodologies are used in its investigation. Quantitative data is more exact and more factual in nature and is subject to less fluctuation. Quantitative tools yield more reliable and objective results, which help in making accurate and effective decisions.

NOTES

Statistical enquiry has become an integral part of practically every decision-making process. The famous science fiction writer, H.G. Wells had considerable foresight when he predicted over a century ago that, 'statistical thinking will one day be as necessary for effective citizenship as the ability to read and write'. Statistical enquiry, indeed, has become an integral part of our informed living where personal and business interactions are being routinely influenced by statistical knowledge and thinking. An average individual is involved in statistical enquiry, knowingly or unknowingly, every day of his life, whether it is comparing prices during shopping or putting an extra lock on his door as a result of reading about the crime rates in the newspapers. The use of statistics has spread to such diverse fields as agriculture, business, economics, medicine, political science, public administration, sociology, psychology, and so on. No field is immune to statistical enquiry.

2.3 NORMAL DISTRIBUTION

Among all the probability distributions the normal probability distribution is by far the most important and frequently used continuous probability distribution. This is so because this distribution well fits in many types of problems. This distribution is of special significance in inferential statistics since it describes probabilistically the link between a statistic and a parameter (*i.e.*, between the sample results and the population from which the sample is drawn). The name of Karl Gauss, eighteenth century mathematician-astronomer, is associated with this distribution and in honour of his contribution, this distribution is often known as the Gaussian distribution.

The normal distribution can be theoretically derived as the limiting form of many discrete distributions. For instance, if in the binomial expansion of $(p + q)^n$, the value of 'n' is infinity and $p = q = \frac{1}{2}$, then a perfectly smooth symmetrical curve would be obtained. Even if the values of p and q are not equal but if the value of the exponent 'n' happens to be very very large, we get a curve normal probability smooth and symmetrical. Such curves are called normal probability curves (or at times known as normal curves of error) and such curves represent the normal distributions.

The probability function in case of normal probability distribution is given as:

$$f(x) = \frac{1}{\sigma.\sqrt{2\pi}} e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where, μ = Mean of the distribution,

σ^2 = Variance of the distribution.

The normal distribution is thus defined by two parameters viz., μ and σ_2. This distribution can be represented graphically as under:

NOTES

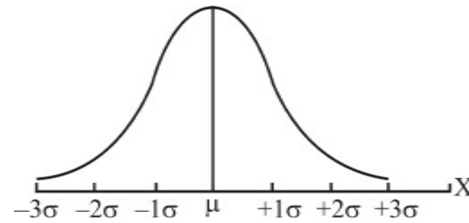
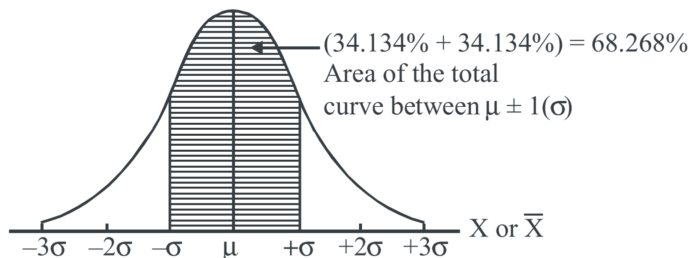


Fig 2.2 *Curve Representing Normal Distribution*

Characteristics of Normal Distribution

The characteristics of the normal distribution or that of normal curve are as given below:

1. It is symmetric distribution.
2. The mean μ defines where the peak of the curve occurs. In other words, the ordinate at the mean is the highest ordinate. The height of the ordinate at a distance of one standard deviation from mean is 60.653% of the height of the mean ordinate and similarly the height of other ordinates at various standard deviations (σ_s) from mean happens to be a fixed relationship with the height of the mean ordinate.
3. The curve is asymptotic to the base line which means that it continues to approach but never touches the horizontal axis.
4. The variance (σ^2) defines the spread of the curve.
5. Area enclosed between mean ordinate and an ordinate at a distance of one standard deviation from the mean is always 34.134% of the total area of the curve. It means that the area enclosed between two ordinates at one sigma (S.D.) distance from the mean on either side would always be 68.268% of the total area. This can be shown as follows:



NOTES

Similarly, the other area relationships are as follows:

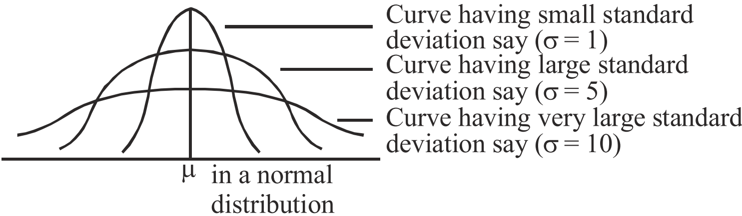
Between		Area covered to total area of the normal curve[4]
$\mu \pm 1$	S.D.	68.27%
$\mu \pm 2$	S.D.	95.45%
$\mu \pm 3$	S.D.	99.73%
$\mu \pm 1.96$	S.D.	95%
$\mu \pm 2.578$	S.D.	99%
$\mu \pm 0.6745$	S.D.	50%

6. The normal distribution has only one mode since the curve has a single peak. In other words, it is always a unimodal distribution.
7. The maximum ordinate divides the graph of normal curve into two equal parts.
8. In addition to all the above stated characteristics the curve has the following properties:
   (*i*) $\mu = \overline{x}$
   (*ii*) $\mu_2 = \sigma^2 =$ variance
   (*iii*) $\mu_4 = 3\sigma^4$
   (*iv*) Moment Coefficient of Kurtosis = 3

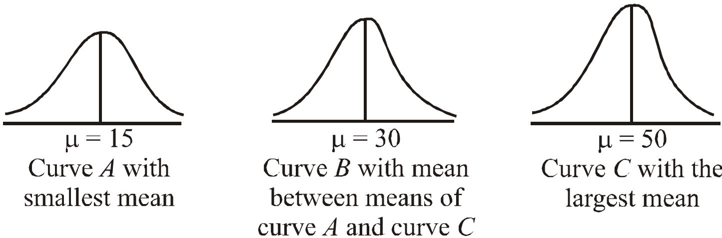
Family of Normal Distributions

We can have several normal probability distributions but each particular normal distribution is being defined by its two parameters viz., the mean (μ) and the standard deviation (σ). There is, thus, not a single normal curve but rather a family of normal curves. We can exhibit some of these as under:

Normal curves with identical means but different standard deviations:



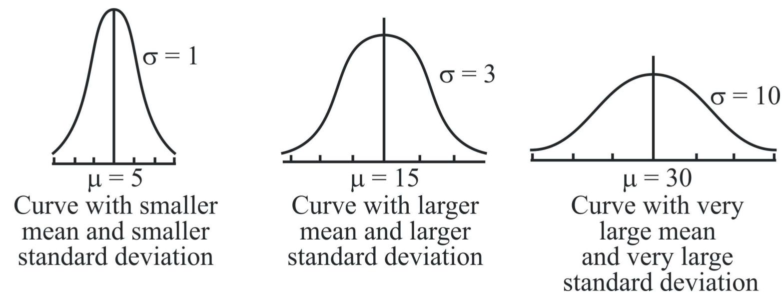
Normal curves with identical standard deviation but each with different means:



Notes:

NOTES

Normal curves each with different standard deviations and different means:



How to measure the area under the Normal Curve?

We have stated above some of the area relationships involving certain intervals of standard deviations (plus and minus) from the means that are true in case of a normal curve. But what should be done in all other cases? We can make use of the statistical tables constructed by mathematicians for the purpose. Using these tables we can find the area (or probability, taking the entire area of the curve as equal to 1) that the normally distributed random variable will lie within certain distances from the mean. These distances are defined in terms of standard deviations. While using the tables showing the area under the normal curve we talk in terms of standard variate (symbolically Z) which really means standard deviations without units of measurement and this 'Z' is worked out as under:

$$Z = \frac{x - \mu}{\sigma}$$

Where, Z = The standard variate (or number of standard deviations from x to the mean of the distribution);

x = Value of the random variable under consideration;

μ = Mean of the distribution of the random variable;

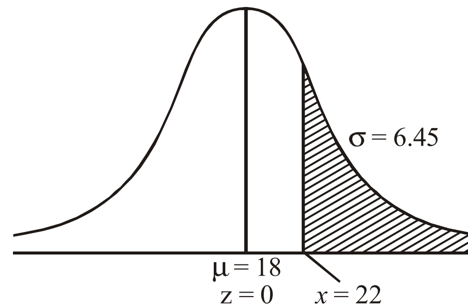
σ = Standard deviation of the distribution.

The table showing the area under the normal curve (often termed as the standard normal probability distribution table) is organized in terms of standard variate (or Z) values. It gives the values for only half the area under the normal curve, beginning with Z = 0 at the mean. Since the normal distribution is perfectly symmetrical the values true for one half of the curve are also true for the other half. We now illustrate the use of such a table for working out certain problems.

Example 2.1: A banker claims that the life of a regular saving account opened with his bank averages 18 months with a standard deviation of 6.45 months. Answer the following: (a) What is the probability that there will still be money in 22 months in a savings account opened with the said bank by a depositor? (b) What is the probability that the account will have been closed before two years?

Solution: (a) For finding the required probability we are interested in the area of the portion of the normal curve as shaded and shown below:

NOTES

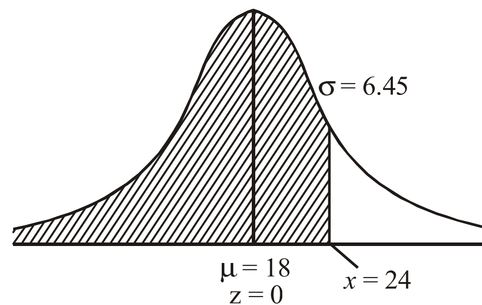


Let us calculate Z as under:

$$Z = \frac{x - \mu}{\sigma} = \frac{22 - 18}{6.45} = 0.62$$

The value from the table showing the area under the normal curve for Z= 0.62 is 0.2324. This means that the area of the curve between μ = 18 and x = 22 is 0.2324. Hence, the area of the shaded portion of the curve is (0.5) – (0.2324) = 0.2676 since the area of the entire right hand portion of the curve always happens to be 0.5. Thus the probability that there will still be money in 22 months in a savings account is 0.2676.

(b) For finding the required probability we are interested in the area of the portion of the normal curve as shaded and shown in figure:



For the purpose we calculate,

$$Z = \frac{24 - 18}{6.45} = 0.93$$

The value from the concerning table, when Z = 0.93, is 0.3238 which refers to the area of the curve between μ = 18 and x = 24. The area of the entire left hand portion of the curve is 0.5 as usual.

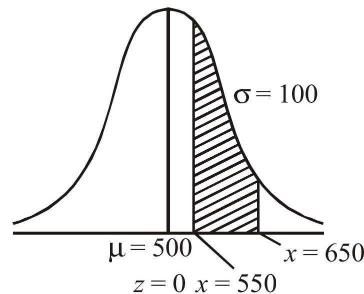
Hence, the area of the shaded portion is (0.5) + (0.3238) = 0.8238 which is the required probability that the account will have been closed before two years, *i.e.*, before 24 months.

NOTES

Example 2.2: Regarding a certain normal distribution concerning the income of the individuals we are given that mean=500 rupees and standard deviation =100 rupees. Find the probability that an individual selected at random will belong to income group,

(a) Rs 550 to Rs 650; (b) Rs 420 to 570.

Solution: (a) For finding the required probability we are interested in the area of the portion of the normal curve as shaded and shown below:



For finding the area of the curve between x = 550 to 650. Let us do the following calculations:

$$Z = \frac{550 - 500}{100} = \frac{50}{100} = 0.50$$

Corresponding to which the area between μ = 500 and x = 550 in the curve as per table is equal to 0.1915 and

$$Z = \frac{650 - 500}{100} = \frac{150}{100} = 1.5$$

Corresponding to which the area between μ = 500 and x = 650 in the curve as per table is equal to 0.4332

Hence, the area of the curve that lies between x = 550 and x = 650 is,

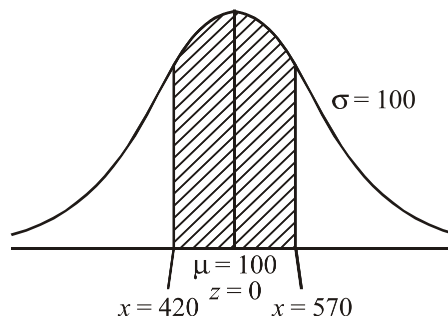
(0.4332) – (0.1915) = 0.2417

This is the required probability that an individual selected at random will belong to income group of Rs 550 to Rs 650.

(b) For finding the required probability we are interested in the area of the portion of the normal curve as shaded and shown below:

NOTES

To find the area of the shaded portion we make the following calculations:



$$Z = \frac{570 - 500}{100} = 0.70$$

Corresponding to which the area between μ = 500 and x = 570 in the curve as per table is equal to 0.2580.

and $$Z = \frac{420 - 500}{100} = -0.80$$

Corresponding to which the area between μ = 500 and x = 420 in the curve as per table is equal to 0.2881.

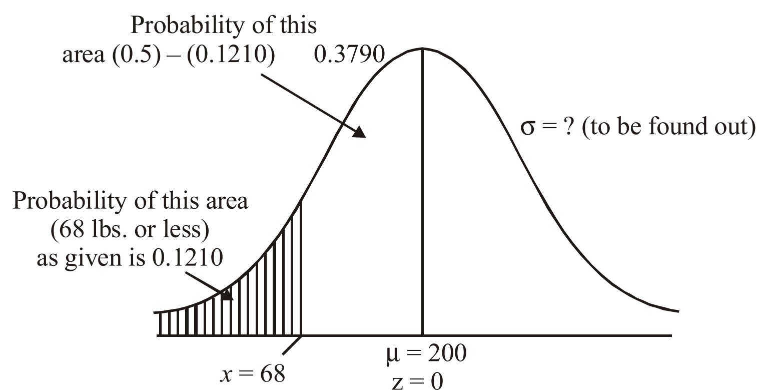
Hence, the required area in the curve between x = 420 and x = 570 is,

(0.2580) + (0.2881) = 0.5461

This is the required probability that an individual selected at random will belong to income group of Rs 420 to Rs 570.

Example 2.3: A certain company manufactures $1\frac{1''}{2}$ all-purpose rope made from imported hemp. The manager of the company knows that the average load-bearing capacity of the rope is 200 lbs. Assuming that normal distribution applies, find the standard deviation of load-bearing capacity for the $1\frac{1''}{2}$ rope if it is given that the rope has a 0.1210 probability of breaking with 68 lbs. or less pull.

Solution: Given information can be depicted in a normal curve as shown below:



NOTES

If the probability of the area falling within $\mu = 200$ and $x = 68$ is 0.3790 as stated above, the corresponding value of Z as per the table[5] showing the area of the normal curve is – 1.17 (minus sign indicates that we are in the left portion of the curve)

Now to find σ we can write,

$$Z = \frac{x - \mu}{\sigma}$$

or $$-1.17 = \frac{68 - 200}{\sigma}$$

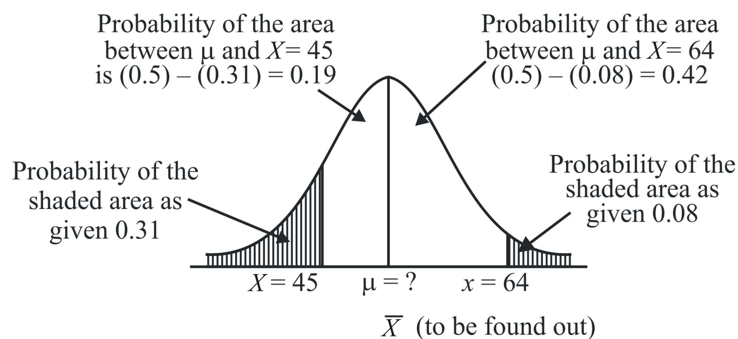
or $$-1.17\sigma = -132$$

or $$\sigma = 112.8 \text{ lbs. approx.}$$

Thus, the required standard deviation is 112.8 lbs. approximately.

Example 2.4: In a normal distribution, 31% items are below 45 and 8% are above 64. Find the $\overline{X}$ and σ of this distribution.

Solution: We can depict the given information in a normal curve as shown below:



If the probability of the area falling within μ and $x = 45$ is 0.19 as stated above, the corresponding value of Z from the table showing the area of the normal curve is – 0.50. Since, we are in the left portion of the curve so we can express this as under,

$$-0.50 = \frac{45 - \mu}{\sigma} \tag{2.1}$$

Similarly, if the probability of the area falling within μ and $x = 64$ is 0.42 as stated above, the corresponding value of Z from the area table is + 1.41. Since, we are in the right portion of the curve so we can express this as under,

$$1.41 = \frac{64 - \mu}{\sigma} \tag{2.2}$$

NOTES

If we solve equations (2.1) and (2.2) above to obtain the value of μ or $\bar{X}$, we have,

$$-0.5\ \sigma = 45 - \mu \qquad (2.3)$$

$$1.41\ \sigma = 64 - \mu \qquad (2.4)$$

By subtracting the equation (2.4) form (2.3) we have,

$$-1.91\ \sigma = -19$$

$$\therefore \qquad \sigma = 10$$

Putting σ = 10 in equation (2.3) we have,

$$-5 = 45 - \mu$$

$$\therefore \qquad \mu = 50$$

Hence, $\bar{x}$ (or μ)=50 and σ =10 for the concerning normal distribution.

2.4 CHI-SQUARE

Chi-square test is a non-parametric test of statistical significance for bivariate tabular analysis (also known as cross-breaks). Any appropriate test of statistical significance lets you know the degree of confidence you can have in accepting or rejecting a hypothesis. Typically, the Chi-square test is any statistical hypothesis test in which the test statistics has a chi-square distribution when the null hypothesis is true. It is performed on different samples (of people) who are different enough in some characteristic or aspect of their behaviour that we can generalize from the samples selected. The population from which our samples are drawn should also be different in the behaviour or characteristic. Amongst the several tests used in statistics for judging the significance of the sampling data, Chi-square test, developed by Prof. Fisher, is considered as an important test. Chi-square, symbolically written as χ^2 (pronounced as Ki-square), is a statistical measure with the help of which, it is possible to assess the significance of the difference between the observed frequencies and the expected frequencies obtained from some hypothetical universe. Chi-square tests enable us to test whether more than two population proportions can be considered equal. In order that Chi-square test may be applicable, both the frequencies must be grouped in the same way and the theoretical distribution must be adjusted to give the same total frequency which is equal to that of observed frequencies. χ^2 is calculated with the help of the following formula:

$$\chi^2 = \sum \left\{ \frac{(f_0 - f_e)^2}{f_e} \right\}$$

Where, f_0 means the observed frequency; and

f_e means the expected frequency.

NOTES

Whether or not a calculated value of χ^2 is significant, it can be ascertained by looking at the tabulated values of χ^2 (given at the end of this book in appendix part) for given degrees of freedom at a certain level of confidence (generally a 5 per cent level is taken). If the calculated value of χ^2 exceeds the table value, the difference between the observed and expected frequencies is taken as significant, but if the table value is more than the calculated value of χ^2, then the difference between the observed and expected frequencies is considered as insignificant, i.e., considered to have arisen as a result of chance and as such can be ignored.

Degrees of Freedom

The number of independent constraints determines the number of degrees of freedom (or df). If there are 10 frequency classes and there is one independent constraint, then there are (10 – 1) = 9 degrees of freedom. Thus, if n is the number of groups and one constraint is placed by making the totals of observed and expected frequencies equal, df = (n – 1); when two constraints are placed by making the totals as well as the arithmetic means equal then df = (n – 2), and so on. In the case of a contingency table (i.e., a table with two columns and more than two rows or table with two rows but more than two columns or a table with more than two rows and more than two columns) or in the case of a 2 × 2 table, the degrees of freedom is worked out as follows:

$$df = (c - 1)(r - 1)$$

Where, c = Number of columns

r = Number of rows

Conditions for the application of test

The following conditions should be satisfied before the test can be applied:

(i) Observations recorded and used are collected on a random basis.

(ii) All the members (or items) in the sample must be independent.

(iii) No group should contain very few items, say less than 10. In cases where the frequencies are less than 10, regrouping is done by combining the frequencies of adjoining groups so that the new frequencies become greater than 10. Some statisticians take this number as 5, but 10 is regarded as better by most of the statisticians.

(iv) The overall number of items (i.e., N) must be reasonably large. It should at least be 50, howsoever small the number of groups may be.

(v) The constraints must be linear. Constraints which involve linear equations in the cell frequencies of a contingency table (i.e., equations containing no squares or higher powers of the frequencies) are known as linear constraints.

Areas of Application of Chi-Square Test

Chi-square test is applicable in large number of problems. The test is, in fact, a technique through the use of which it is possible for us to:

(a) Test the Goodness of Fit;

(b) Test the Homogeneity of a Number of Frequency Distributions;

(c) Test the Significance of Association between Two Attributes

NOTES

In other words, Chi-square test is a test of independence, goodness of fit and homogeneity. At times Chi-square test is used as a test of population variance also.

As a test of goodness of fit, χ^2 test enables us to see how well the distribution of observed data fits the assumed theoretical distribution, such as Binomial distribution, Poisson distribution or the Normal distribution.

As a test of independence, χ^2 test helps explain whether or not two attributes are associated. For instance, we may be interested in knowing whether a new medicine is effective in controlling fever or not and χ^2 test will help us in deciding this issue. In such a situation, we proceed on the null hypothesis that the two attributes (viz., new medicine and control of fever) are independent. Which means that the new medicine is not effective in controlling fever. It may, however, be stated here that χ^2 is not a measure of the degree of relationship or the form of relationship between two attributes but it simply is a technique of judging the significance of such association or relationship between two attributes.

As a test of homogeneity, χ^2 test helps us in stating whether different samples come from the same universe. Through this test, we can also explain whether the results worked out on the basis of sample/samples are in conformity with well-defined hypothesis or the results fail to support the given hypothesis. As such, the test can be taken as an important decision-making technique.

As a test of population variance. Chi-square is also used to test the significance of population variance through confidence intervals, especially in case of small samples.

So far, our estimate of population parameters or comparison of sample and population characteristics have involved certain assumptions about the populations from which the samples have been drawn. In most statistical tests, we have based our decisions on the assumption that the population was normally distributed. Even in the case of the binomial distribution, we approximated it to a normal distribution so that Z score test could be used to make certain decisions.

When this assumption about the population cannot be made, then it becomes necessary to use other procedures. One of the tests used in such situations is known as Chi Square *(χ^2)* Test. This test is good for *nominal* or *ordinal* scale of measurement, where nominal scale of measurement deals with the data which can only be classified into categories such as male and female, or freshman, juniors and seniors, and so on. There is no particular order for these groupings and furthermore, all categories are separate and mutually exclusive so that an item in one category is not included in another category. The ordinal level of measurement assigns different ranks to these categories. One category may be superior in standing and the other category may be good or fair and so on.

NOTES

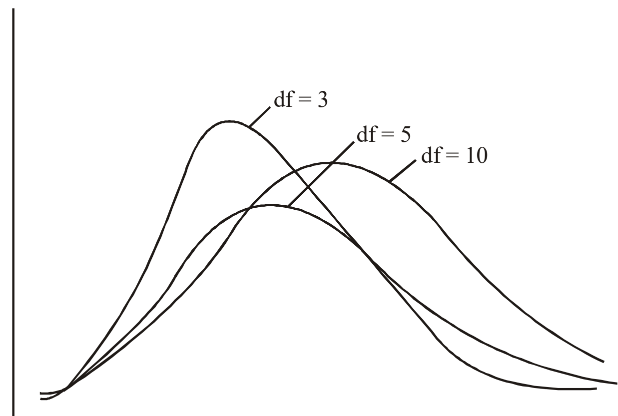
Similar to the binomial distribution, χ^2 test is also used for analysing qualitative variables such as opinions of persons, religious affiliation, smoking habits, and so on. However, unlike binomial distribution test which deals with comparison of two population proportions, the χ^2 test deals with judgements about proportions of two or more than two populations.

χ^2 distribution has the following properties.

- It involves squared observations and hence, it is always positive. Its value is always greater than or equal to zero.
- The distribution is not symmetrical. It is skewed to the right so that its skewness is positive. However, as the number of degrees of freedom increases, **Chi Square** approaches a symmetric distribution.
- Similar to *t*-distribution, there is a family of Chi Square distributions. There is a particular distribution for each degree of freedom.

The estimation of degrees of freedom for χ^2 distribution is different than such estimation for *t*-distribution. While in *t*-distribution, the degrees of freedom are determined by the sample size as (n–1), for χ^2 distribution, these are determined by the number of categories in which various attributes of the sample are placed so that if there are (k) number of categories, then the number of degrees of freedom would be (k–1). For example, if a sample of 100 students were categorized as freshman, sophomores, juniors and seniors, then there would be 4 categories so that k = 4, and (k – 1) = 3 degrees of freedom.

The following illustration shows the family of χ^2 curves with varying degrees of freedom and it can be seen that as the number of degrees of freedom increases, χ^2 distribution approaches the normal curve.



The χ^2 test is used to test whether there is a significant difference between the *observed* number of responses in each category and the *expected* number of responses for such category under the assumptions of null hypothesis. In other words, the objective is to find out how well the distribution of observed frequencies (f_o) fit the distribution of expected frequencies (f_e). Hence, this test is also called *goodness-of-fit* test.

NOTES

This test can best be illustrated with an example.

Example 2.5: Suppose that we take a sample of 200 businessmen from New York, out of which 100 believe that the economic conditions would improve, then we take a sample of 300 businessmen from Washington, out of which 100 believe that the economic conditions would improve and we take a sample of 250 businessmen from Chicago and 120 of them believe that economic conditions would improve. Now, we want to test whether there is a significant difference between the opinions of the businessmen from these three different cities at a given level of significance.

The null hypothesis will assume that there is no difference between opinions of businessmen in these three cities, which can be considered as categories. The alternative hypothesis would be that all these categories are not similar.

The random variable whose sampling distribution is approximated by χ^2 distribution is given by:

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

where

f_o = observed frequency of responses in a given category

f_e = expected frequency of responses in the same category under the assumption of the null hypothesis

The calculated value of χ^2 is then compared with the critical value of χ^2 from the table with a pre-established value at the level of significance α and the appropriate value of the degrees of freedom. The degrees of freedom (df) are calculated as follows:

a) df = (k – 1), where k is the number of categories in one-sample test (example follows).

b) df = (k – l)(r – 1), where k is the number of columns and *r* is the number of rows in a cross-classification table (known as contingency table) for various categories of two or more independent samples.

(It should be noted that the critical value separates the region of acceptance from the region of rejection.)

As stated earlier, a χ^2 distribution is a family of distributions with a different distribution for each value of (df). The degrees of freedom can be defined as the number of observations that are free to vary after the required restrictions are placed on the data. For example, if a data with 30 responses are placed in two categories such as favour or do not favour an issue, and 20 of these are known to fall in one category, then we know that the balance of 10 responses must fall in the other category. In this case, the number of degrees of freedom for two categories (k) is (k – 1) = 1, since only the number of responses in one category is free to

vary, while the responses in the other category are fixed by the total number of responses and the number of responses in the first category.

χ^2 One-Sample Test

NOTES

Assume that a die was rolled 30 times to check if the die was fair or loaded. If it is a fair and balanced die, then we should expect each face to come up five times since the probability (p) of each face of fair die coming up is 1/6 and the expected value of each face coming up in 30 rolls is = np = (30 × 1/6) = 5.

In the experiment conducted, the actual number of times each face came up in sequence is as follows:

4, 7, 3, 6, 8, 2

A comparision of observed frequency and expected frequency of each face coming up in 30 rolls is tabulated as follows.

	Face Value					
	1	2	3	4	5	6
Observed frequency	4	7	3	6	8	2
Expected frequency	5	5	5	5	5	5

Steps involved in the Process

Step 1. State the null hypothesis and the alternate hypothesis.

H_o: All faces are equally likely to occur. In other words,

$$p_1 = p_2 = p_3 = p_4 = p_5 = p_6$$

H_1: All probabilities are not equal or at least two of the probabilities (or proportions) differ from each other.

Under null hypothesis, all proportions must be equal. Even if one of these proportions is not equal to any of the others, the null hypothesis cannot be accepted.

Step 2. A level of significance is selected.

Assume $\alpha = 0\ 05$. This is the probability of making type I error. This means that when $\alpha = 0.05$, we will be making an error of rejecting the null hypothesis when in fact it is true, 5 per cent of the times.

Step 3. Calculate the expected frequency f_e for each category.

In our case, $f_e = 5$.

Step 4. Use an appropriate test statistic.

In our case, χ^2 test is selected because we are comparing observed frequencies with expected frequencies in discrete categories. (The categories are the six faces of the die) χ^2 test measures the discrepancy between the observed values and the expected values for decision-making purposes about the null hypothesis, so that:

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

NOTES

In our case,

$$\chi^2 = \frac{(4-5)^2}{5} + \frac{(7-5)^2}{5} + \frac{(3-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(8-5)^2}{5} + \frac{(2-5)^2}{5}$$

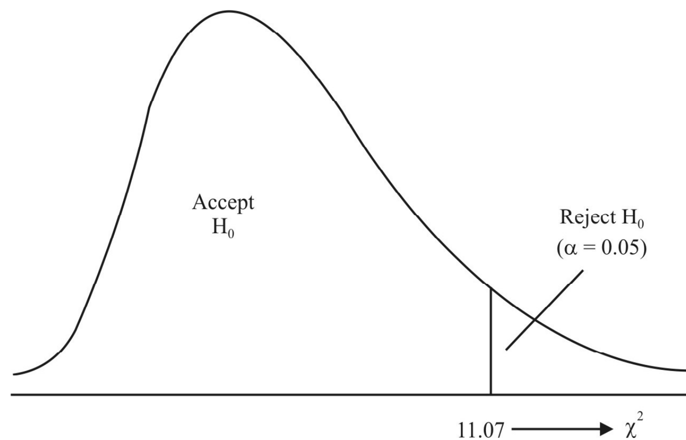
$$= \frac{1}{5} + \frac{4}{5} + \frac{4}{5} + \frac{1}{5} + \frac{9}{5} + \frac{9}{5}$$

$$= \frac{28}{5} = 5.6$$

Step 5. A decision rule is formulated.

We check the critical value of χ^2 from the table against $\alpha = 0.05$ and df = (k – 1) = (6 – 1) = 5. This value is given as 11.07. We compare our computed value of χ^2 with the critical value of χ^2 from the table. Since our computed value of $\chi^2 = 5.6$ is less than the critical value of $\chi^2 = 11.07$, we cannot reject the null hypothesis.

The following diagram of χ^2 distribution illustrates this point.



The one sample of χ^2 test is restricted by the following observations:

1. When k = 2 and df = (k – 1) = 1, then each expected frequency should be at least 5.
2. When df > 1, then χ^2 for one-sample test should not be used if more than 20 per cent of the expected frequencies are smaller than 5 in value or when any expected frequency is smaller than 1.

Example 2.6: Suppose that 60 children were asked as to which ice-cream flavour they liked out of the three flavours of vanilla, strawberry and chocolate. The answers are recorded as follows:

Flavour	Number
Vanilla	17
Strawberry	24
Chocolate	19

Our objective is to determine whether children favour any particular flavour compared to other flavours.

Solution: The null hypothesis states that there is no difference among the tastes of children as far as the ice-cream flavours are concerned. Under the null hypothesis, equal number of children are expected to prefer each flavour. This means that the expected frequencies should be 20 for vanilla, 20 for strawberry and 20 for chocolate.

The table of expected frequencies and observed frequencies is shown as follows:

Flavour	*Observed Frequency*	*Expected Frequency*
Vanilla	17	20
Strawberry	24	20
Chocolate	19	20

Using the χ^2 test, we get,

$$\chi^2 = \Sigma\frac{\left(f_o - f_e\right)^2}{f_e}$$

$$\chi^2 = \frac{(17-20)^2}{20}+\frac{(24-20)^2}{20}+\frac{(19-20)^2}{20}$$

$$= \frac{9}{20}+\frac{16}{20}+\frac{1}{20}$$

$$= \frac{26}{20} = 1.3$$

If we assume the level of significance $\alpha = 0.05$ and knowing the degrees of freedom df = (k –1), where k is the number of categories which in our case is 3 so that df = (3 – 1) = 2, then we can compare our computed value of χ^2 with critical value of χ^2 from the table at $\alpha = 0.05$ and df = 2 and reach a decision whether to accept or reject the null hypothesis.

The critical value of χ^2 is given as = 5.991. Since our computed value of χ^2 is less than the critical value of χ^2, we cannot reject the null hypothesis.

χ^2 Test-Contingency Tables

In the previous section, we discussed χ^2 test for goodness-of-fit for a single trait only. However, we may be confronted with a situation where we want to determine the significance of differences in characteristics between two or more groups. For example, we may want to test if there are any significant differences between males and females in adjustment to old age where adjustment can be classified into categories of good, fair, average, poor, and so on. The data concerning the relative frequencies with which the group members fall into various categories, is

then presented in the form of a table consisting of rows and columns and the format is known as the contingency table. The rows and columns are used to summarize and display the results of data collected and are categorized on the basis of classification of categories.

NOTES

Testing Hypothesis for Independence of Two Categories

Although χ^2 test is used as goodness-of-fit test, it is most often used as a test of independence to determine if the paired observations obtained on two or more nominal variables are independent of each other or not. It is sometimes necessary to deal with the idea of two variables being related to one another in the sense that the value of one variable depends upon the value of the other corresponding variable. For example, if we were testing for the degree of association between the height and the weight of persons, it would be easy to recognize that tall people are expected to weigh more than the short people. Hence, the variables height and weight are not independent of each other. Similarly, the variables of income and education seem to be related to each other. However, the age and the colour of the eyes are not related to each other, so that knowing one would not influence the knowledge about the other. Similarly, there may or may not be any association between the opinion on nuclear disarmament and the sex of the person. Such dependence or independence can be tested by χ^2 test.

In testing for any relationship between the opinion on nuclear disarmament and the sex of the person, let us assume that 100 persons including 60 males were asked about their opinion and their responses were classified into two categories *yes* and *no* as follows:

	Male	Female	Total
Yes	35	25	60
No	25	15	40
Total	60	40	100

Now, if the opinions were totally independent of the sex then the same proportion of males and females would favour nuclear disarmament. This means that men would be no more likely to favour nuclear disarmament than women, and vice versa.

This table shows that 60 out of 100 (or 60%) persons surveyed favoured nuclear disarmament. Under the null hypothesis, if the opinions are independent of sex, then 60% of males as well as 60% of females are expected to favour disarmament. This means:

60% of males = 60% of 60 = 36

Similarly, 60% of females = 60% of 40 = 24

These expected frequencies are calculated as follows:

Let us call the small rectangles with observed frequencies as *cells.* Then the expected frequency for each *cell* can be calculated by multiplying the total of the row containing that cell with the total of the column containing that cell and dividing this product by the grand total of the table.

NOTES

For example, if we wanted to know the expected frequency of the first cell in the North-West corner of the table representing males who favour disarmament, then the total of the row is 60 and the total of that column is 60 and the grand total of the table is 100, and hence,

$$\text{Expected frequency for that cell} = \frac{60 \times 60}{100} = 36$$

Thus, the frequency of each cell can be calculated. These expected frequencies are then written on the right hand corner (or any corner) in a small rectangle to identify these as expected frequencies against the observed frequencies which are written in the middle of the cell as follows:

	Male	Female	
Yes	35 [36]	25 [24]	60
No	25 [24]	15 [16]	40
Total	60	40	100

This must be clear that the sum of the expected frequencies in each category must be the same as the sum of the observed frequencies.

Then the null hypothesis for independence can be tested by:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where:

O_{ij} = Observed frequency in the cell identified by the intersection of the ith row and jth column

E_{ij} = Expected frequency in the cell identified by the intersection of the ith row and jth column

r = Number of rows

k = Number of columns

$\sum_{i=1}^{r}\sum_{j=1}^{k}$ = Summation of all cells in all rows (r) and all column (k).

Simply written and in accordance with the symbols used inthe previous section, χ^2 can be written as:

$$\chi^2 = \Sigma\frac{(f_o - f_e)^2}{f_e}$$

NOTES

where

f_o = Observed frequency

f_e = Expected frequency

Σ = Summation over all cells

The degrees of freedom in a contingency table are given as:

$$df = (r - 1)(k - 1)$$

where

r = number of rows

k = number of columns

Solving this problem we get the value of χ^2 as,

$$= \frac{(35-36)^2}{36} + \frac{(25-24)^2}{24} + \frac{(25-24)^2}{24} + \frac{(15-16)^2}{16}$$

$$= \frac{1}{36} + \frac{1}{24} + \frac{1}{24} + \frac{1}{16}$$

$$= 0.028 + 0.041 + 0.041 + 0.062$$

$$= 0.172$$

Looking at the value of χ^2 from the table for 95% confidence (or $\alpha = 0.05$) and df = (2 – 1)(2 – 1) = 1, we get,

$$\chi^2 = 3.841$$

Since our calculated value of χ^2 is less than the critical value of $\chi^2 = 3.841$, we cannot reject the null hypothesis that the opinion is independent of sex.

Example 2.7: A study was conducted among 100 professors from 3 different divisions at a community college for their promotions based upon 3 categories of teaching, research and other college activities. It is required to test if there is any relationship between the basis for promotion and the field of teaching. The observed values of the number of professors promoted in each category are given as follows in the contingency table.

Field of Teaching

Basis for Promotion	Business	Science	Nursing	Total
Teaching	20 [16.0]	10 [11.2]	10 [12.8]	40
Research	10 [14.0]	10 [9.8]	15 [11.2]	35
Other	10 [10.0]	8 [7.0]	7 [8.0]	25
Total	40	28	32	100

NOTES

Solution.

1. State the null hypothesis. The null hypothesis states that there is no association between the basis for promotion and the field of teaching. The alternate hypothesis would be that there is an association between the field of teaching and the basis for promotion, so that professors in some categories are more likely to be promoted than in other categories.
2. Construct a two-way contingency table (as shown here) with the observed values in each respective cell.
3. Establish a level of significance. Let $\alpha = 0.05$ in our case.
4. Determine the expected value for each cell, under the assumption of null hypothesis, by using the following formula:

The expected frequency of a given cell = Total of the row of the cell multiplied by the total of the column of the cell and divided by the grand total.

These expected frequencies are then recorded alongside the respective observed frequencies in a small rectangle on the corner of each cell. The expected frequencies for each cell thus are computed as follows:

Cell 1. Teaching/Business $\dfrac{40 \times 40}{100}$ $= 16.0$

Cell 2. Teaching/Science $\dfrac{40 \times 28}{100}$ $= 11.2$

Cell 3. Teaching/Nursing $\dfrac{40 \times 32}{100}$ $= 12.8$

Cell 4. Research/Business $\dfrac{35 \times 40}{100}$ $= 14.0$

Cell 5. Research/Science $\dfrac{35 \times 28}{100}$ $= 9.8$

Cell 6. Research/Nursing $\dfrac{35 \times 32}{100}$ $= 11.2$

Cell 7. Other/Business $\dfrac{25 \times 40}{100}$ $= 10.0$

Cell 8. Other/Science $\dfrac{25 \times 28}{100}$ $= 7.0$

Cell 9. Other/Nursing $\dfrac{25 \times 32}{100}$ $= 8.0$

NOTES

5. Compute the value of χ^2 by using the formula:

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(20-16)^2}{16} + \frac{(10-11.2)^2}{11.2} + \frac{(10-12.8)}{12.8} + \frac{(10-14)^2}{14} + \frac{(10-9.8)^2}{9.8} + \frac{(15-11.2)^2}{11.2}$$

$$+ \frac{(10-10)^2}{10} + \frac{(8-7)^2}{7} + \frac{(7-8)^2}{8}$$

= 1 + 0.128 + 0.612 + 1.14 + 0.004 + 1.289 + 0 + 0.143 + 0.125

= 4.441

6. Check the critical value of χ^2 from the table for $\alpha = 0.05$ and $df = (r-1)(k-1) = (3-1)(3-1) = 4$, and we get the value 9.488.
7. By comparsion of the two values of χ^2, we see that our calculated value of χ^2 is less than the critical value of χ^2. We cannot reject the null hypothesis.

Example 2.8: 200 digits are chosen at random from a set of tables. The frequencies of the digits are as follow:

Digit	0	1	2	3	4	5	5	7	8	9
Frequency	18	19	23	21	16	25	22	20	21	15

Use χ^2 test to assess the correctness of the hypothesis that the digits were distributed in equal numbers in the tables from they were chosen.

Solution: H_o = the digits were distributed in equal numbers.

F_o	F_e	$(F_o - F_e)^2$	$\left(\frac{F_o - F_e}{F_e}\right)^2$
18	20	4	0.20
19	20	1	0.05
23	20	9	0.45
21	20	1	0.05
16	20	16	0.80
25	20	25	1.25
22	20	4	0.20
20	20	0	0.00
21	20	1	0.05
15	20	25	1.25
200	200		4.30 χ^2

Expected frequency (F_e) $\frac{200}{10}$=20

$\chi^2 = 4.3$

$df\,(k - 1) = 10 - 1 = 9$

$\chi^2_{0.05} = 16.22$

The calculated value of χ^2 4.3 is less than the critical value of χ^2 16.22 from the table. Hence the hypothesis that the digits were distributed in equal numbers is accepted.

Example 2.9: The demand for particular spare part in a factory was found to vary from day to day. In a sample study the following information was obtained.

Days	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Demand	1124	1125	1110	1120	1126	1115

Test the hypothesis that the number of parts demanded does not depend on the day of week, the table value of χ^2 for 5 d.f at 5% level of significance is 11.07.

Solution: H_o = that the number of parts demanded does not depend on the day of week.

Day	F_o	F_e	$(F_o - F_e)^2$	$\left(\frac{F_o - F_e}{F_e}\right)^2$
Monday	1124	1120	16	0.014
Tuesday	1125	1120	25	0.022
Wednesday	1110	1120	100	0.089
Thursday	1120	1120	0	0.000
Friday	1126	1120	36	0.032
Saturday	1115	1120	25	0.022
	6720			0.179

$$\text{Expected frequency} = \frac{1124+1125+1110+1120+1126+1115}{6}$$

$$= \frac{6720}{6} = 1120$$

$\chi^2_{0.05} = 11.07$

$\chi^2 = 0.179$

$179 < 11.07$

The calculated value of χ^2 is less the critical value of the table hence H_o is true the number of parts demanded does not depend on the day of week.

NOTES

Example 2.10: Based on information on 500 randomly selected field about the tennacy status of the cultivators of these fields and use of fertilizers collected in an agro-economic enquiry the following classification was noted . Can one conclude that owner-cultivators are more inclined towards the use of fertilizers?

	Owned	Rented
Using fertilizer	208	92
Not using fertilizer	32	168

Solution: H_o = Owner cultivators are not more inclined towards the use of fertilizers.

F_o

208	92	300
32	168	200
240	260	500

F_e

144	156	300
96	104	200
240	260	500

F_o	F_e	$(F_o - F_e)^2$	$\left(\frac{F_o - F_e}{F_e}\right)^2$
208	144	4096	28.44
32	96	4096	42.67
92	156	4096	26.26
168	104	4096	39.38
			136.75

$df = 1$

χ^2 0.05 = 3.84

3.84 < 136.75

Example 2.11: 1000 families were selected at random in a city to test the belief that high income families usually send their children to public school and low income families often sent their children to government schools.

The following results were obtained.

Income	Public School	Government School	Total
Low	370	430	800
High	130	70	200
Total	500	500	1000

Test whether income and type of schooling are independent.

Solution: H_o is that the income and type of schooling are independent.

F_o

370	430	800
130	70	200
500	500	1000

F_e

400	400	800
100	100	200
500	500	1000

NOTES

χ^2 test.

F_o	F_e	$(F_o - F_e)^2$	$\left(\frac{F_o - F_e}{F_e}\right)^2$
370	400	900	2.25
130	100	900	9.00
430	400	900	2.25
70	100	900	9.00
			22.50

Var K. $(r-1)(c-1)(2-1)(2+1) = 1$

$V_1 = \chi^2_{0.05}$ 3.84

22.5 > 3.84

The calculated value of χ^2 is more then the table values the H_0 is rejected. Hence income and type of schooling are not independent.

Example 2.12: In a survey of 200 boys of which 75 were intelligent 40 had educated fathers. 85 of the unintelligent boys had uneducated fathers. So these figures support the hypothesis that educated fathers have intelligent boys? Use χ^2-test, value of χ^2 for 1 degree of freedom at 5% level is 3.84.

Given data are:

F_o

	Educated Father	Uneducated Father	Total
Intelligent boys	40	35	75
Unintelligent boys	40	85	125
Total	80	120	200

F_e

30	45	75
50	75	125
80	120	200

Solution: The hypothesis that there is no association between the educated father in son.

F_o	F_e	$(F_o - F_e)^2$	$\left(\frac{F_o - F_e}{F_e}\right)^2$
40	30	100	3.33
40	50	100	2.00
35	45	100	2.22
85	75	100	1.33
			8.88

$$\chi^2 = 8.88$$

$$V = 1 = \chi^2_{0.05} = 3.84$$

NOTES

The calculated value of χ^2 is higher than table value. The hypothesis is rejected. Hence educated fathers have intelligent boys.

Example 2.13: A sample analysis of examination result of 450 students was made. It was found that 220 had failed, 120 had secured third class. 90 were placed in second class and 20 got first class. Are these figures commensurate with the general examination result which is in the ratio of 4:3:2:1 for various categories respectively?

	Failed	III	II	I
F_o =	220	120	90	20

$F_e = 220 + 120 + 90 – 20 = 450$

Ratio = 4 + 3 + 2 + 1 = 10

Failed $\frac{4}{1\not{0}} \times 45\not{0} = 180$

III $\frac{3}{1\not{0}} \times 45\not{0} = 135$

II $\frac{2}{1\not{0}} \times 45\not{0} = 90$

I $\frac{1}{1\not{0}} \times 45\not{0} = 45$

F_o	F_e	$(F_o - F_e)^2$	$\left(\frac{F_o + F_e}{F_e}\right)^2$
220	180	1600	8.888
120	135	225	1.667
90	90	0	0.000
20	45	625	13.889
			24.444

Since the calculated value of χ^2 24.44 is greater than (24.44 > 7.81) the table value our hypothesis does not hold good.

2.5 *t*-DISTRIBUTIONS

Sir William S. Gosset (pen name Student) developed a significance test and through it made significant contribution to the theory of sampling applicable in case of small samples. When population variance is not known, the test is commonly known as Student's *t*-test and is based on the *t*-distribution.

NOTES

Like the normal distribution, *t*-distribution is also symmetrical but happens to be flatter than the normal distribution. Moreover, there is a different *t*-distribution for every possible sample size. As the sample size gets larger, the shape of the *t*-distribution loses its flatness and becomes approximately equal to the normal distribution. In fact, for sample sizes of more than 30, the *t*-distribution is so close to the normal distribution that we will use the normal to approximate the *t*-distribution. Thus, when *n* is small, the *t*-distribution is far from normal, but when *n* is infinite, it is identical to normal distribution.

For applying *t*-test in context of small samples, the *t* value is calculated first of all and, then the calculated value is compared with the table value of *t* at certain level of significance for given degrees of freedom. If the calculated value of *t* exceeds the table value (say $t_{0.05}$), we infer that the difference is significant at 5 per cent level, but if the calculated value is t_0, is less than its concerning table value, the difference is not treated as significant.

The t-test is used when the following two conditions are fullfiled:

(i) The sample size is less than 30, i.e., when $n \leq 30$.

(ii) The population standard deviation (σ_p) must be unknown.

In using the *t*-test, we assume the following:

(i) The population is normal or approximately normal.

(ii) The observations are independent and the samples are randomly drawn samples.

(iii) There is no measurement error.

(iv) In the case of two samples, population variances are regarded as equal if equality of the two population means is to be tested.

The following formulae are commonly used to calculate the *t* value:

(i) To test the significance of the mean of a random sample

$$t = \frac{|\bar{X} - \mu|}{S \,|\, SE_{\bar{x}}\bar{X}}$$

Where, $\bar{X}$ = Mean of the sample

μ = Mean of the universe

$SE_{\bar{x}}$ = S.E. of mean in case of small sample and is worked out as,

$$SE_{\bar{x}} = \frac{\sigma_s}{\sqrt{n}} = \frac{\sqrt{\frac{\Sigma(x_i - \bar{x})^2}{\sqrt{n}}}}{\sqrt{n}}$$

and the degrees of freedom = $(n - 1)$

NOTES

The above stated formula for t can as well be stated as,

$$t = \frac{|\bar{x} - \mu|}{SE_{\bar{x}}}$$

$$= \frac{|\bar{x} - \mu|}{\frac{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}}{\sqrt{n}}}$$

$$= \frac{|\bar{x} - \mu|}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}} \times \sqrt{n}$$

If we want to work out the probable or fiducial limits of population mean (μ) in case of small samples, we can use either of the following:

(*a*) Probable limits with 95 per cent confidence level:

$$\mu = \bar{X} \pm SE_{\bar{x}} \ (t_{0.05})$$

(*b*) Probable limits with 99 per cent confidence level:

$$\mu = \bar{X} \pm SE_{\bar{x}} \ (t_{0.01})$$

At other confidence levels, the limits can be worked out in a similar manner, taking the concerning table value of t just as we have taken $t_{0.05}$ in (a) and $t_{0.01}$ in (b) above.

(ii) To test the difference between the means of two samples

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\text{SE}_{\bar{x}_1 - \bar{x}_2}}$$

Where, $\bar{X}_1$ = Mean of the sample 1

$\bar{X}_2$ = Mean of the sample 2

$SE_{\bar{x}_1 - \bar{x}_2}$ = Standard error of difference between two sample means and is worked out as follows:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sum(X_{1i} - \bar{x}_1)^2 + \sum(X_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and the degrees of freedom = $(n_1 + n_2 - 2)$.

When the actual means are in fraction, then use of assumed means is convenient. In such a case, the standard deviation of difference, i.e.,

$$\sqrt{\frac{\Sigma(x_{1i}+x_1)^2+\Sigma(x_{2i}-\bar{x}_2)^2}{n_1+n_2-2}}$$

can be worked out by the following short-cut formula:

$$=\frac{\sqrt{\Sigma(x_{1i}-A_1)^2}+\Sigma(x_{2i}-A_1)^2-n_1(x_{1i}-A_2)^2-n_2(x_{2i}-A_2)^2}{n_1+n_2-2}$$

Where, A_1 = Assumed mean of sample 1

A_2 = Assumed mean of sample 2

X_1 = True mean of sample 1

X_2 = True mean of sample 2

(iii) To test the significance of an observed correlation coefficient

$$t=\frac{r}{\sqrt{1-r^2}}\times\sqrt{n-2}$$

Here, t is based on $(n-2)$ degrees of freedom.

(iv) In context of the 'difference test'

Difference test is applied in the case of paired data and in this context t is calculated as,

$$t=\frac{\bar{x}_{Ditt}-0}{6_{Diff}\sqrt{n}}=\frac{\bar{x}_{Diff}-0}{6_{Diff}}\sqrt{n}$$

Where, $\bar{X}_{Diff}$ or $\bar{D}$ = Mean of the differences of sample items.

0 = the value zero on the hypothesis that there is no difference

$\sigma_{Diff.}$ = standard deviation of difference and is worked out as

$$\sqrt{\frac{\sum D-\bar{X}_{Diff})^2}{(n-1)}}$$

or

$$\sqrt{\frac{\Sigma D^2-(\bar{D})^2 n}{(n-1)}}$$

D = differences

n = number of pairs in two samples and is based on $(n-1)$ degrees of freedom

NOTES

Check Your Progress

1. Define descriptive statistics.
2. Elaborate on the inferential statistics.
3. What is statistical enquiry?
4. Explain about the normal distribution.
5. What do you understand by Chi-square?
6. Give the properties of χ^2 distribution.
7. Define the *t*-test.

2.6 *F*-DISTRIBUTIONS

In business decisions, we are often involved in determining if there are significant differences among various sample means, from which conclusions can be drawn about the differences among various population means. What if we have to compare more than two sample means? For example, we may be interested to find out if there are any significant differences in the average sales figures of four different salesmen employed by the same company, or we may be interested to find out if the average monthly expenditures of a family of 4 in 5 different localities are similar or not, or the telephone company may be interested in checking, whether there are any significant differences in the average number of requests for information received in a given day among the five areas of New York City, and so on. The methodology used for such types of determinations is known as Analysis of Variance.

This technique is one of the most powerful techniques in statistical analysis and was developed by R.A. Fisher. It is also called the *F*-Test.

There are two types of classifications involved in the analysis of variance. The one-way analysis of variance refers to the situations when only one fact or variable is considered. For example, in testing for differences in sales for three salesman, we are considering only one factor, which is the salesman's selling ability. In the second type of classification, the response variable of interest may be affected by more than one factor. For example, the sales may be affected not only by the salesman's selling ability, but also by the price charged or the extent of advertising in a given area.

For the sake of simplicity and necessity, our discussion will be limited to One-way Analysis of Variance (ANOVA).

The null hypothesis, that we are going to test, is based upon the assumption that there is no significant difference among the means of different populations. For example, if we are testing for differences in the means of k populations, then,

$$H_0 = \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$$

The alternate hypothesis (H_1) will state that at least two means are different from each other. In order to accept the null hypothesis, all means must be equal. Even if one mean is not equal to the others, then we cannot accept the null hypothesis. The simultaneous comparison of several population means is called *Analysis of Variance or ANOVA*.

NOTES

Assumptions

The methodology of ANOVA is based on the following assumptions.

(i) Each sample of size n is drawn randomly and each sample is independent of the other samples.

(ii) The populations are normally distributed.

(iii) The populations from which the samples are drawn have equal variances. This means that:

$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \text{.........} = \sigma_k^2,$ for k populations.

The rationale behind analysis of variance

Why do we call it the Analysis of Variance, even though we are testing for means? Why not simply call it the Analysis of Means? How do we test for means by analysing the variances? As a matter of fact, in order to determine if the means of several populations are equal, we do consider the measure of variance, σ^2.

The estimate of population variance, σ^2, is computed by two different estimates of σ^2, each one by a different method. One approach is to compute an estimator of σ^2 in such a manner that even if the population means are not equal, it will have no effect on the value of this estimator. This means that, the differences in the values of the population means do not alter the value of σ^2 as calculated by a given method. This estimator of σ^2 is the average of the variances found within each of the samples. For example, if we take 10 samples of size *n*, then each sample will have a mean and a variance. Then, the mean of these 10 variances would be considered as an unbiased estimator of σ^2, the population variance, and its value remains appropriate irrespective of whether the population means are equal or not. This is really done by pooling all the sample variances to estimate a common population variance, which is the average of all sample variances. This common variance is known as variance within samples or σ^2_{within}.

The second approach to calculate the estimate of σ^2, is based upon the Central Limit Theorem and is valid only under the null hypothesis assumption that all the population means are equal. This means that in fact, if there are *no differences* among the population means, then the computed value of σ^2 by the second approach should not differ significantly from the computed value of σ^2 by the first approach.

Hence,

If these two values of σ^2 are approximately the same, then we can decide to accept the null hypothesis.

NOTES

The second approach results in the following computation:

Based upon the Central Limit Theorem, we have previously found that the standard error of the sample means is calculated by,

$$\sigma^2_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

or, the variance would be:

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

or, $\sigma^2 = n\sigma^2_{\bar{x}}$

Thus, by knowing the square of the standard error of the mean $(\sigma_{\bar{x}})^2$, we could multiply it by n and obtain a precise estimate of σ^2. This approach of estimating σ^2 is known as $\sigma^2_{between}$. Now, if the null hypothesis is true, that is if all population means are equal then, $\sigma^2_{between}$ value should be approximately the same as σ^2_{within} value. A significant difference between these two values would lead us to conclude that this difference is the result of differences between the population means.

But, how do we know that any difference between these two values is significant or not? How do we know whether this difference, if any, is simply due to random sampling error or due to actual differences among the population means?

R.A. Fisher developed a Fisher test or F-test to answer the above question. He determined that the difference between $\sigma^2_{between}$ and σ^2_{within} values could be expressed as a ratio to be designated as the F-value, so that,

$$F = \frac{\sigma^2_{between}}{\sigma^2_{within}}$$

In the minters case, if the population means are exactly the same, then $\sigma^2_{between}$ will be equal to the σ^2_{within} and the value of F will be equal to 1.

However, because of sampling errors and other variations, some disparity between these two values will be there, even when the null hypothesis is true, meaning that all population means are equal. The extent of disparity between the two variances and consequently, the value of F, will influence our decision on whether to accept or reject the null hypothesis. It is logical to conclude that, if the population means are not equal, then their sample means will also vary greatly from one another, resulting in a larger value of $\sigma^2_{between}$ and hence a larger value of F (σ^2_{within} is based only on sample variances and not on sample means and hence, is not affected by differences in sample means). Accordingly, the larger the value of F, the more likely the decision to reject the null hypothesis. But, how large the value of F be so as to reject the null hypothesis? The answer is that the computed value of F must

be larger than the *critical* value of F, given in the table for a given level of significance and calculated number of degrees of freedom. (The F-distribution is a family of curves, so that there are different curves for different degrees of freedom).

Degrees of freedom

We have talked about the F-distribution being a family of curves, each curve reflecting the degrees of freedom relative to both $\sigma^2_{between}$ and σ^2_{within}. This means that, the degrees of freedom are associated both with the numerator as well as with the denominator of the F-ratio.

(i) **The numerator.** Since the variance between samples, $s^2_{between}$ comes from many samples and if there are k number of samples, then the degrees of freedom, associated with the numerator would be $(k-1)$.

(ii) **The denominator** is the *mean variance* of the variances of k samples and since, each variance in each sample is associated with the size of the sample (n), then the degrees of freedom associated with each sample would be $(n - 1)$. Hence, the total degrees of freedom would be the sum of the degrees of freedom of k samples or

$df = k(n-1)$, when each sample is of size n.

The *F*-distribution

The major characteristics of the F-distribution are as follows:

(*i*) Unlike normal distribution, which is only one type of curve irrespective of the value of the mean and the standard deviation, the F distribution is a *family* of curves. A particular curve is determined by two parameters. These are the degrees of freedom in the numerator and the degrees of freedom in the denominator. The shape of the curve changes as the number of degrees of freedom changes.

(*ii*) It is a continuous distribution and the value of F cannot be negative.

(*iii*) The curve representing the F distribution is positively skewed.

(*iv*) The values of F theoretically range from zero to infinity.

A diagram of F distribution curve is shown in Figure 2.3.

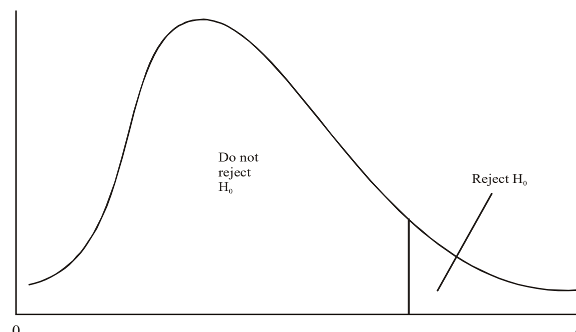


Fig. 2.3 *F-Distribution on Curve*

NOTES

The rejection region is only in the right end tail of the curve because unlike Z distribution and t-distribution which had negative values for areas below the mean, F-distribution has only positive values by definition and only positive values of F that are larger than the critical values of F, will lead to a decision to reject the null hypothesis.

Computation of *F*

F-ratio contains only two elements, which are the variance between the samples and the variance within the samples.

If all the means of samples were exactly equal and all samples were exactly representative of their respective populations so that all the sample means were exactly equal to each other and to the population mean, then there will be no variance. However, this can never be the case. We always have variation, both between samples and within samples, even if we take these samples randomly and from the same population. This variation is known as the total variation.

The total variation designated by $\sum(X - \bar{\bar{X}})^2$, where X represents individual observations for all samples and $\bar{\bar{X}}$ is the grand mean of all sample means and equals (μ), the population mean, is also known as the *total sum of squares* or *SST,* and is simply the sum of squared differences between each observation and the overall mean. This total variation represents the contribution of two elements. These elements are:

(i) Variance between samples: The variance between samples may be due to the effect of different *treatments,* meaning that the population means may be affected by the *factor* under consideration, thus making the population means actually different, and some variance may be due to the inter-sample variability. This variance is also known as the sum of squares between samples. Let this sum of squares be designated as *SSB*.

Then, *SSB* is calculated by the following steps:

a. Take k samples of size n each and calculate the mean of each sample, i.e., $\bar{X}_1, \bar{X}_2, \bar{X}_3, ... \bar{X}_k$.

b. Calculate the grand mean $\bar{\bar{X}}$ of the distribution of these sample means, so that,

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{k} \bar{x}_i}{k}$$

c. Take the difference between the means of the various samples and the grand mean, i.e.,

$$(\bar{X}_1 - \bar{\bar{X}}),\ (\bar{X}_2 - \bar{\bar{X}}),\ (\bar{X}_3 - \bar{\bar{X}}),\ ....,\ (\bar{X}_k - \bar{\bar{X}})$$

d. Square these deviations or differences individually, multiply each of these squared deviations by its respective sample size and sum up all these products, so that we get;

$$\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2, \text{ where } n_i = \text{size of the } i\text{th sample.}$$

NOTES

This will be the value of the *SSB*.

However, if the individual observations of all samples are not available, and only the various means of these samples are available, where the samples are either of the same size *n* or different sizes, $n_i, n_2, n_3, \ldots, n_k$, then the value of SSB can be calculated as:

$$SSB = n_i(\bar{X}_i - \bar{\bar{X}})^2 + n_2(\bar{X}_2 - \bar{\bar{X}})^2 + \ \ldots.. \ n_k(\bar{X}_k - \bar{\bar{X}})^2$$

Where,

n_1 = number of items in sample 1

n_2 = number of items in sample 2

n_k = number of items in sample *k*

$\bar{X}_1$ = mean of sample 1

$\bar{X}_2$ = mean of sample 2

$\bar{X}_k$ = mean of sample *k*

$\bar{\bar{X}}$ = Grand mean or average of all items in all samples.

e. Divide *SSB* by the degrees of freedom, which are $(k - 1)$, where *k* is the number of samples and this would give us the value of $\sigma^2_{\text{between}}$, so that,

$$\sigma^2_{\text{between}} = \frac{SSB}{(k-1)}.$$

(This is also known as mean square between samples or *MSB)*.

(ii) Variance within samples: Even though each observation in a given sample comes from the same population and is subjected to the same treatment, some chance variation can still occur. This variance may be due to sampling errors or other natural causes. This variance or sum of squares is calculated by the following steps:

a. Calculate the mean value of each sample, i.e., $\bar{X}_1, \bar{X}_2, \bar{X}_3, \ \ldots. \ \bar{X}_k$.

b. Take one sample at a time and take the deviation of each item in the sample from its mean. Do this for all the samples, so that we would have a difference between each value in each sample and their respective means for all values in all samples.

c. Square these differences and take the total of all these squared differences (or deviations). This sum is also known as *SSW* or sum of squares within samples.

NOTES

d. Divide this *SSW* by the corresponding degrees of freedom. The degrees of freedom are obtained by subtracting the total number of samples from the total number of items. Thus, if N is the total number of items or observations, and k is the number of samples, then,

$$df = (N - k)$$

These are the degrees of freedom within samples. (If all samples are of equal size n, then $df = k(n-1)$, since $(n-1)$ are the degrees of freedom for each sample and there are k samples).

e. This figure *SSW*/*df*, is also known as σ^2_{within}, or MSW (mean of sum of squares within samples).

Now, the value of F can be computed as:

$$F = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{SSB/df}{SSW/df}$$

$$= \frac{SSB/(k-1)}{SSW/(N-k)} = \frac{MSB}{MSW}$$

This value of F is then compared with the critical value of F from the table and a decision is made about the validity of null hypothesis.

2.7 ESTIMATION OF PARAMETERS

Parameter estimates (also called coefficients) are the change in the response associated with a one-unit change of the predictor, all other predictors being held constant. The unknown model parameters are estimated using least-squares estimation. Parameter estimation is a branch of statistics that which leads to the using sample data to estimate the parameters of a distribution.

Methods of Parameter Estimation

The techniques used for parameter estimation are called estimators.

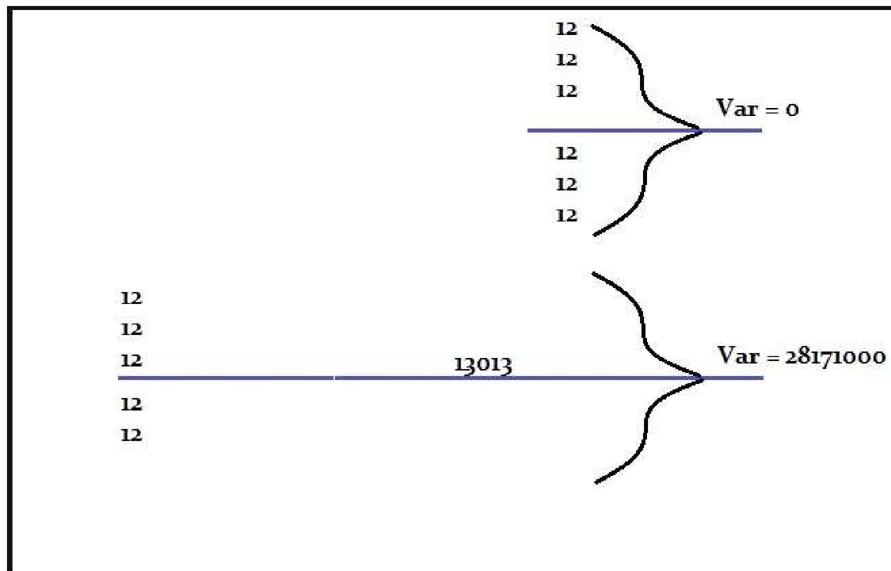
Some estimators are:

- **Probability Plotting:** A method of finding parameter values where the data is plotted on special plotting paper and parameters are derived from the visual plot.
- **Rank Regression (Least Squares):** A method of finding parameter values that minimizes the sum of the squares of the residuals.
- **Maximum Likelihood Estimation:** A method of finding parameter values that, given a set of observations, will maximize the likelihood function.
- **Bayesian Estimation Methods:** A family of estimation methods that tries to minimize the posterior expectation of what is called the utility

function. In practice, what this means is that existing knowledge about a situation is formulated, data is gathered, and then posterior knowledge is used to update our beliefs.

Qualities of Estimators

If the value an estimator estimates for the parameter, θ', always converges to the actual parameter value θ as the quantity of data used for parameter estimation increases, we say an estimator is consistent. The bias of an estimator is the deviation of the expectation from the actual true value. If, for a given estimator, the bias is zero, we say that that estimator is unbiased.



Parameters are descriptive measures of an entire population that may be used as the inputs for a Probability Distribution Function (PDF) to generate distribution curves. Parameters are usually signified by Greek letters to distinguish them from sample statistics. For example, the population mean is represented by the Greek letter mu (μ) and the population standard deviation by the Greek letter sigma (σ). Parameters are fixed constants, that is, they do not vary like variables. However, their values are usually unknown because it is infeasible to measure an entire population.

Each distribution is entirely defined by several specific parameters, usually between one and three. The following table provides examples of the parameters required for three distributions. The parameter values determine the location and shape of the curve on the plot of distribution, and each unique combination of parameter values produces a unique distribution curve.

Distribution	Parameter 1	Parameter 2	Parameter 3
Chi-square	Degrees of freedom		
Normal	Mean	Standard deviation	
3-Parameter Gamma	Shape	Scale	Threshold

NOTES

Parameters are descriptive measures of an entire population. However, their values are usually unknown because it is infeasible to measure an entire population. Because of this, you can take a random sample from the population to obtain parameter estimates. One goal of statistical analyses is to obtain estimates of the population parameters along with the amount of error associated with these estimates. These estimates are also known as sample statistics.

There are Several Types of Parameter Estimates:

- Point estimates are the single, most likely value of a parameter. For example, the point estimate of population mean (the parameter) is the sample mean (the parameter estimate).
- Confidence intervals are a range of values likely to contain the population parameter.

For an example of parameter estimates, suppose you work for a spark plug manufacturer that is studying a problem in their spark plug gap. It would be too costly to measure every single spark plug that is made. Instead, you randomly sample 100 spark plugs and measure the gap in millimetres. The mean of the sample is 9.2. This is the point estimate for the population mean (μ). You also create a 95% confidence interval for μ which is (8.8, 9.6). This means that you can be 95% confident that the true value of the average gap for all the spark plugs is between 8.8 and 9.6.

2.8 PROPERTIES OF ESTIMATORS

A point estimate uses a single sample value to estimate the desired population parameter. For example, a sample mean $\bar{x}$ is considered as a point estimate of the population mean μ. Similarly, a sample standard deviation s is a point estimate of population standard deviation σ. For instance, if we want to know the grade point average (GPA) of seniors majoring in Business Administration at Medgar Evers College, then we take a random sample of business major seniors and calculate the sample mean $\bar{x}$ of the sample. Then, the value of this $\bar{x}$ would be considered as a point estimate of μ which is the grade point average of the entire population of students majoring in business administration. Similarly, the sample variance s^2 is the point estimate of the population variance σ^2.

In point estimate, we seek the sample statistic, such as $\bar{x}$, computed from sample observations, which is the best estimate of the corresponding population parameter, such as μ. But how do we know that the sample statistic that we computed from sample observations is the best estimator of the population parameter? By *best* we mean that the value of the sample statistic should be as close to the population parameter as possible. For example, if the sample mean grade point average for business students is calculated as 3.5 out of 4, then the population average grade point average should also be 3.5 or very close to it in order for sample average to be a good estimator of population average. Since the

NOTES

population parameter is always inferred from sample statistic, it is necessary and important that such sample statistic should be as highly reliable as an estimator for population parameter, as possible. For example, there are three measures of central tendency, namely, mean, mode and median for a sample that can be used as point estimators for the population average. It is important to know as to which one of these measures best represents the population mean. As an illustration, suppose that we want to find out the average time that a salesman for a company spends with the customer. Suppose further that we took a sample and found out that on an average, a salesman spent 60 minutes with a customer (mean). However, most salesmen spent 45 minutes (the mode) and the median was 65 minutes. The question now is to establish as to which of these measures would best describe the population parameter as to how much time on an average a salesman spends with the customer?

Properties of Good Estimate

The best estimator should be highly reliable and have such desirable properties as unbiasedness, consistency, efficiency and sufficiency. These criteria are described as follows:

(a) ***Unbiasedness.*** An estimator is a random variable since it is always a function of the sample values. For example, the value of the sample average would depend upon the values of the sample and may differ from sample to sample. The expected value of the sample average is considered to be an unbiased estimator if it equals the population mean which is being estimated. This means that:

$$E\ (\bar{x}) = \mu$$

(Since sampling distribution is a probability distribution, we refer to the average as expected value instead of simply the average).

(b) ***Consistency.*** Consistency refers to the effect of sample size on the accuracy of the estimator. A statistic is said to be consistent estimator of the population parameter, if it approaches the parameter as the sample size increases, so that in the case of the mean:

$$\bar{x} \to \mu \text{ as } n \to N$$

(c) ***Efficiency.*** An estimator is considered to be efficient if its value remains stable from sample to sample. The best estimator would be the one which would have the least variance from sample to sample taken randomly from the same population. From the three point estimators of central tendency, namely, the mean, the mode and the median, the mean is considered to be the least variant and hence a better estimator.

(d) ***Sufficiency.*** An estimator is said to be sufficient if it uses all the information about the population parameter contained in the sample. For example, the statistic mean uses all the sample values in its computation while mode and the median do not. Hence the mean is a better estimator in this sense.

NOTES

Some of the parameters of the population and their estimators are as follows:

$$\mu = \bar{x} = \frac{\sum x}{n}$$

$$\sigma = s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

$$p = p_s = \left(\frac{x}{n}\right), \text{ where } p_s \text{ is the sample proportion.}$$

2.9 TESTING OF HYPOTHESES

A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. Hypothesis refers to a provisional idea whose merit needs evaluation, but having no specific meaning. Though it is often referred to as a convenient mathematical approach for simplifying cumbersome calculation. Setting up and testing hypotheses is an integral art of statistical inference. Hypotheses are often statements about population parameters like variance and expected value. During the course of hypothesis testing, some inference about population like the mean and proportion are made. Any useful hypothesis will enable predictions by reasoning including deductive reasoning. According to Karl Popper a hypothesis must be falsifiable and that a proposition or theory cannot be called scientific if it does not admit the possibility of being shown false. Hypothesis might predict outcome of an experiment in a lab setting the observation of a phenomenon in nature. Thus, hypothesis is a explanation of a phenomenon proposal suggesting a possible correlation between multiple phenomena.

The characteristics of hypothesis are:

- **Clear and accurate**: Hypothesis should be clear and accurate so as to draw a consistent conclusion.
- **Statement of relationship between variables**: If a hypothesis is relational, it should state the relationship between different variables.
- **Testability**: A hypothesis should be open to testing so that other deductions can be made from it and can be confirmed or disproved by observation. The researcher should do some prior study to make the hypothesis a testable one.
- **Specific with limited scope**: A hypothesis, which is specific, with limited scope, is easily testable than a hypothesis with limitless scope. Therefore, a researcher should pay more time to do research on such kind of hypotheses.
- **Simplicity**: A hypothesis should be stated in the most simple and clear terms to make it understandable.
- **Consistency**: A hypothesis should be reliable and consistent with established and known facts.

NOTES

- **Time-Limit**: A hypothesis should be capable of being tested within a reasonable time. In other words, it can be said that the excellence of a hypothesis is judged by the time taken to collect the data needed for the test.
- **Empirical reference**: A hypothesis should explain or support all the sufficient facts needed to understand what the problem is all about.

A hypothesis is a statement or assumption concerning a population. For the purpose of decision-making, a hypothesis has to be verified and then accepted or rejected. This is done with the help of observations. We test a sample and make a decision on the basis of the result obtained. Decision-making plays significant role in different areas such as marketing, industry and management.

Statistical Decision-Making

Testing a statistical hypothesis on the basis of a sample enables us to decide whether the hypothesis should be accepted or rejected. The sample data enable us to accept or reject the hypothesis. Since the sample data give incomplete information about the population the result of the test need not be considered to be final or unchallengeable. The procedure, which, on the basis of sample results, enables us to decide whether a hypothesis is to be accepted or rejected, is called Hypothesis Testing or Test of Significance.

Note **:** A test provides evidence, if any, against a hypothesis, usually called a mill hypothesis. The test cannot prove the hypothesis to be correct. It can give some evidence against it.

The hypothesis makes some assumption about the density function of the random variate. The sampling distribution is fundamental to this subject.

The test of a hypothesis means a procedure to decide whether to accept or reject a hypothesis.

If a sample is found to have an untenable probability (of occurrence) level (called the significance level), we reject the hypothesis. Usually the probability levels of 0.05 and 0.01 are taken. They are called 5% and 1% significance levels.

Note**:** The acceptance of a hypothesis implies there is no evidence from the sample that we should believe otherwise.

The rejection of a hypothesis leads us to conclude that it is false. This way of putting the problem is convenient because of the uncertainty inherent in the problem. In view of this we must always briefly state a hypothesis that we hope to reject.

A hypothesis stated in the hope of being rejected is called a null hypothesis and is denoted by H_0.

If H_0 is rejected, it may lead to the acceptance of an alternative hypothesis denoted

by H_1.

For example, new fragrance soap is introduced in the market. The null hypothesis H_0, which may be rejected, is that the new soap is not better than the existing soap.

NOTES

Example 2.14: A die is suspected to be loaded. Roll the die a number of times to test.

Solution: The null hypothesis H_0: $p = 1/6$ for showing six.

The alternative hupothesis H_1: $p \neq 1/6$

Null and Alternative Hypothesis

Hypothesis is usually considered as the principal instrument in research. The basic concepts regarding the testability of a hypothesis are as follows:

(a) **Null Hypothesis and Alternative Hypothesis**: In the context of statistical analysis, while comparing any two methods, the following concepts or assumptions are taken into consideration:

(*i*) **Null Hypothesis**: While comparing two different methods in terms of their superiority, wherein the assumption is that both the methods are equally good is called null hypothesis. It is also known as statistical hypothesis and is symbolised as $H_{0.}$

(*ii*) **Alternate Hypothesis**: While comparing two different methods, regarding their superiority, wherein, stating a particular method to be good or bad as compared to the other one is called alternate hypothesis. It is symbolised as H_0.

(b) **Comparison of Null Hypothesis with Alternate Hypothesis**: Following are the points of comparison between null hypothesis and alternate hypothesis:

(*i*) Null hypothesis is always specific while Alternate Hypothesis gives an approximate value.

(*ii*) The rejection of Null hypothesis involves great risk, which is not in the case of Alternate hypothesis.

Null hypothesis is more frequently used in statistics than Alternate hypothesis because it is specific and is not based on probabilities.

The hypothesis to be tested is called the Null Hypothesis and is denoted by H_0. This is to be tested against other possible states of nature called alternative hypotheses. The alternative is usually denoted by $H_{1.}$

The null hypothesis implies that there is no difference between the statistic and the population parameter. To test whether there is no difference between the sample mean x and the population μ, we write the null hypothesis.

$$H_0\text{: } \bar{x} = \mu$$

NOTES

The alternative hypothesis would be

$$H_0:\ \bar{x} \neq \mu$$

This means $\bar{x} > \mu$ or $\bar{x} < \mu$. This is called a two-tailed hypothesis.

The alternative H_0: $\bar{x} > \mu$ is right tailed.

The alternative H_0: $\bar{x} < \mu$ is left tailed.

These are one sided or one-tailed alternatives.

Note:

1. The alternative hypothesis H_1 implies all such values of the parameter, which are not specified by the null hypothesis H_0.
2. Testing a statistical hypothesis is a rule, which leads to a decision to accept or reject a hypothesis.

A one tailed test requires rejection of the null hypothesis when the sample statistic is greater than the population value or less than the population value at a certain level of significance.

1. We may want to test if the sample mean x exceeds the population mean p. Then the null hypothesis is,

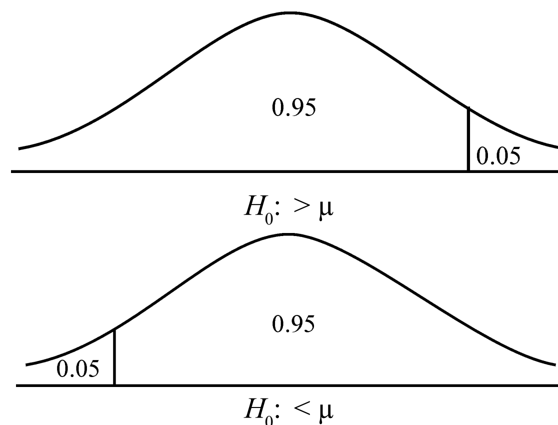
$$H_0:\ \bar{x} > \mu$$

2. In the other case the null hypothesis could be

$$H_0:\ \bar{x} < \mu$$

Each of these two situations leads to a one tailed test and has to be dealt with in the same manner as the two tailed test. Here the critical rejection is on one side only, right for $\bar{x} > \mu$ and left for $\bar{x} < \mu$. Each diagram here shows a five per cent level of test of significance.

For example, a minister in a certain government has an average life of 11 months without being involved in a scam. A new party claims to provide ministers with an average life of more than 11 months without scam. We would like to test if; on the average the new ministers last longer than 11 months. We may write the null hypothesis H_0: $\bar{x} = 11$ and alternative hypothesis H_1: $\bar{x} > 11$ or H_1: $\bar{x} < 11$.



NOTES

Check Your Progress

8. Who developed F-test?
9. Give the assumptions of ANOVA.
10. What is parameter estimation?
11. Explain about the Bayesian estimation methods of parameter.
12. What are the properties of good estimate?
13. Elaborate on the hypotheses testing.
14. What do you understand by null hypothesis?

2.10 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Descriptive statistics describes the data and consists of methods and techniques used in collection, organization, presentation and analysis of data in order to describe the various features and characteristics of such data. These methods can either be graphical or computational.
2. Inferential statistics can be defined as those methods that are used to estimate a characteristic of a population or making a decision concerning a population on the basis of the results obtained from a sample taken from the same population.
3. Statistical enquiry refers to an investigation of a given phenomenon on the basis of statistical and quantitative models and techniques. An enquiry is defined as a close examination of a matter in a search of information or truth. 'Close' here means intense and comprehensive. The matter must be looked in depth and not merely at the surface.
4. Normal distribution is of special significance in inferential statistics since it describes probabilistically the link between a statistic and a parameter (i.e., between the sample results and the population from which the sample is drawn). The name of Karl Gauss, eighteenth century mathematician-astronomer, is associated with this distribution and in honour of his contribution, this distribution is often known as the Gaussian distribution. The normal distribution can be theoretically derived as the limiting form of many discrete distributions.
5. Chi-square test is a non-parametric test of statistical significance for bivariate tabular analysis (also known as cross-breaks). Any appropriate test of statistical significance lets you know the degree of confidence you can have in accepting or rejecting a hypothesis. Typically, the Chi-square test is any statistical hypothesis test in which the test statistics has a chi-square distribution when the null hypothesis is true. Chi-square, symbolically written

as χ^2 (pronounced as Ki-square), is a statistical measure with the help of which, it is possible to assess the significance of the difference between the observed frequencies and the expected frequencies obtained from some hypothetical universe.

6. χ^2 distribution has the following properties.
   - It involves squared observations and hence, it is always positive. Its value is lways greater than or equal to zero.
   - The distribution is not symmetrical. It is skewed to the right so that its skewness is positive. However, as the number of degrees of freedom increases, Chi-square approaches a symmetric distribution.
   - Similar to *t*-distribution, there is a family of Chi-square distributions. There is a particular distribution for each degree of freedom.
7. Sir William S. Gosset (pen name Student) developed a significance test and through it made significant contribution to the theory of sampling applicable in case of small samples. When population variance is not known, the test is commonly known as Student's *t*-test and is based on the *t*-distribution. Like the normal distribution, *t*-distribution is also symmetrical but happens to be flatter than the normal distribution.
8. The *F*-test technique is one of the most powerful techniques in statistical analysis and was developed by R.A. Fisher. It is also called the *F*-Test.
9. The methodology of ANOVA is based on the following assumptions.
   (i) Each sample of size n is drawn randomly and each sample is independent of the other samples.
   (ii) The populations are normally distributed.
   (iii) The populations from which the samples are drawn have equal variances.

   This means that– $\sigma^2_1 = \sigma^2_2 = \sigma^2_3 = ...= \sigma^2_k$, for k populations.
10. Parameter estimates (also called coefficients) are the change in the response associated with a one-unit change of the predictor, all other predictors being held constant. The unknown model parameters are estimated using least-squares estimation. Parameter Estimation is a branch of statistics that which leads to the using sample data to estimate the parameters of a distribution.
11. A family of estimation methods that tries to minimize the posterior expectation of what is called the utility function. In practice, what this means is that existing knowledge about a situation is formulated, data is gathered, and then posterior knowledge is used to update our beliefs.
12. The best estimator should be highly reliable and have such desirable properties as unbiasedness, consistency, efficiency and sufficiency.
13. A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. Hypothesis refers to a provisional

idea whose merit needs evaluation, but having no specific meaning. Though it is often referred to as a convenient mathematical approach for simplifying cumbersome calculation. Setting up and testing hypotheses is an integral part of statistical inference.

NOTES

14. While comparing two different methods in terms of their superiority, wherein the assumption is that both the methods are equally good is called null hypothesis. It is also known as statistical hypothesis and is symbolised as H_0.

2.11 SUMMARY

- As the name suggests, descriptive statistics merely describes the data and consists of methods and techniques used in collection, organization, presentation and analysis of data in order to describe the various features and characteristics of such data.
- Inferential statistics can be defined as those methods that are used to estimate a characteristic of a population or making a decision concerning a population on the basis of results obtained from a sample taken from the same population. The measured characteristics of the sample are known as sample statistics, while the measured characteristics of the population are known as population parameters.
- Statistical enquiry involves objective analysis of information in order to arrive at some meaningful conclusion. For example, the Federal Drug Administration (FDA) approves a drug for consumption after lengthy statistical investigations regarding the results of the experiments about the need and usefulness of the drug.
- Non-statistical enquiry, on the other hand, is gathering of opinions and feelings, and the conclusions so arrived at are more subjective in nature. It is more of an observation rather than an enquiry. Observations and statements about beauty, goodness, honesty, and so on are all non-statistical enquiries.
- Among all the probability distributions the normal probability distribution is by far the most important and frequently used continuous probability distribution. This is so because this distribution well fits in many types of problems. This distribution is of special significance in inferential statistics since it describes probabilistically the link between a statistic and a parameter (i.e., between the sample results and the population from which the sample is drawn).
- The mean μ defines where the peak of the curve occurs. In other words, the ordinate at the mean is the highest ordinate. The height of the ordinate at a distance of one standard deviation from mean is 60.653% of the height of the mean ordinate and similarly the height of other ordinates at various standard deviations (σ_s) from mean happens to be a fixed relationship with the height of the mean ordinate.

NOTES

- Chi-square test is a non-parametric test of statistical significance for bivariate tabular analysis (also known as cross-breaks). Any appropriate test of statistical significance lets you know the degree of confidence you can have in accepting or rejecting a hypothesis. Typically, the Chi-square test is any statistical hypothesis test in which the test statistics has a chi-square distribution when the null hypothesis is true.
- Chi-square, symbolically written as χ^2 (pronounced as Ki-square), is a statistical measure with the help of which, it is possible to assess the significance of the difference between the observed frequencies and the expected frequencies obtained from some hypothetical universe.
- Chi-square test is a test of independence, goodness of fit and homogeneity. At times Chi-square test is used as a test of population variance also. As a test of goodness of fit, χ^2 test enables us to see how well the distribution of observed data fits the assumed theoretical distribution such as Binomial distribution, Poisson distribution or the Normal distribution.
- As a test of homogeneity, χ^2 test helps us in stating whether different samples come from the same universe. Through this test, we can also explain whether the results worked out on the basis of sample/samples are in conformity with well-defined hypothesis or the results fail to support the given hypothesis.
- This test is good for nominal or ordinal scale of measurement, where nominal scale of measurement deals with the data which can only be classified into categories such as male and female, or freshman, juniors and seniors, and so on. There is no particular order for these groupings and furthermore, all categories are separate and mutually exclusive so that an item in one category is not included in another category.
- Similar to the binomial distribution, χ^2 test is also used for analysing qualitative Variables, such as opinions of persons, religious affiliation, smoking habits, and so on. However, unlike binomial distribution test which deals with comparison of two population proportions, the χ^2 test deals with judgements about proportions of two or more than two populations.
- Sir William S. Gosset (pen name Student) developed a significance test and through it made significant contribution to the theory of sampling applicable in case of small samples. When population variance is not known, the test is commonly known as Student's *t*-test and is based on the *t*-distribution. Like the normal distribution, *t*-distribution is also symmetrical but happens to be flatter than the normal distribution.
- Like the normal distribution, t distribution is also symmetrical but happens to be flatter than the normal distribution. Moreover, there is a different *t*-distribution for every possible sample size. As the sample size gets larger, the shape of the *t*-distribution loses its flatness and becomes approximately equal to the normal distribution.

NOTES

- There are two types of classifications involved in the analysis of variance. The one-way analysis of variance refers to the situations when only one fact or variable is considered. For example, in testing for differences in sales for three salesman, we are considering only one factor, which is the salesman's selling ability. In the second type of classification, the response variable of interest may be affected by more than one factor.
- The estimate of population variance, σ^2, is computed by two different estimates of σ^2, each one by a different method. One approach is to compute an estimator of σ^2 in such a manner that even if the population means are not equal, it will have no effect on the value of this estimator.
- The second approach to calculate the estimate of σ^2, is based upon the Central Limit Theorem and is valid only under the null hypothesis assumption that all the population means are equal. This means that in fact, if there are no differences among the population means, then the computed value of σ^2 by the second approach should not differ significantly from the computed value of σ^2 by the first approach.
- The denominator is the mean variance of the variances of k samples and since, each variance in each sample is associated with the size of the sample(n), then the degrees of freedom associated with each sample would be $(n-1)$.
- The variance between samples may be due to the effect of different treatments, meaning that the population means may be affected by the factor under consideration, thus making the population means actually different, and some variance may be due to the inter-sample variability.
- Parameter estimates (also called coefficients) are the change in the response associated with a one-unit change of the predictor, all other predictors being held constant. The unknown model parameters are estimated using least-squares estimation. Parameter Estimation is a branch of statistics that which leads to the using sample data to estimate the parameters of a distribution.
- An estimator is a random variable since it is always a function of the sample values. For example, the value of the sample average would depend upon the values of the sample and may differ from sample to sample. The expected value of the sample average is considered to be an unbiased estimator if it equals the population mean which is being estimated.
- A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. Hypothesis refers to a provisional idea whose merit needs evaluation, but having no specific meaning.
- Testing a statistical hypothesis on the basis of a sample enables us to decide whether the hypothesis should be accepted or rejected. The sample data enable us to accept or reject the hypothesis.

2.12 KEY WORDS

NOTES

- **Descriptive statistics:** Descriptive statistics merely describes the data and consists of methods and techniques used in collection, organization, presentation and analysis of data in order to describe the various features and characteristics of such data.
- **Chi-square:** Chi-square, symbolically written as χ^2 (pronounced as Ki-square), is a statistical measure with the help of which, it is possible to assess the significance of the difference between the observed frequencies and the expected frequencies obtained from some hypothetical universe.
- **Parameter estimation:** Parameter estimation is a branch of statistics that which leads to the using sample data to estimate the parameters of a distribution.
- **Probability Plotting:** A method of finding parameter values where the data is plotted on special plotting paper and parameters are derived from the visual plot.
- **Efficiency:** An estimator is considered to be efficient if its value remains stable from sample to sample. The best estimator would be the one which would have the least variance from sample to sample taken randomly from the same population.
- **Testability:** A hypothesis should be open to testing so that other deductions can be made from it and can be confirmed or disproved by observation. The researcher should do some prior study to make the hypothesis a testable one.

2.13 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Define the term descriptive statistics.
2. Explain the inferential statistics.
3. What is normal distribution?
4. Give the characteristics of normal distribution.
5. What is Chi-square?
6. What do you understand by the χ^2 one-sample test?
7. Elaborate on the t-test.
8. Derive the test of random sample.
9. Interpret the F-distribution.
10. Comprehend the degree of freedom.
11. What is parameter estimation?

NOTES

12. Explain the types of parameter estimates.
13. What do you understand by properties of estimators?
14. Elaborate on the properties of good estimate.
15. Interpret the testing of hypotheses.
16. Give the characteristics of hypothesis.
17. Distinguish between the null and alternative hypothesis.

Long-Answer Questions

1. Briefly explain about the descriptive statistics and statistical enquiry with the help of examples.
2. Describe the inferential statistics with its significance.
3. Explain in detail about the normal distribution and how to measure the area under the normal curve explain with various types of examples.
4. Discuss in detail about the Chi-square with appropriate examples.
5. Analyse the *t*-distribution giving the various test based on it.
6. Discuss in detail about the *F*-distribution with the help of examples.
7. Describe the estimation of parameter with its types.
8. Briefly explain about the estimation of parameters with its types.
9. Analyse the testing of hypotheses with the help of examples.

2.14 FURTHER READINGS

Johnston, J. and John DiNARDO. 1997. *Econometric Methods*, Fourth Edition. New Delhi: Tata McGraw-Hill.

Koutsoyiannis, A.1977. *Theory of Econometrics,* Second Edition. London: The Macmillan Press Ltd.

Özdemir, Durmuº. 2016. *Applied Statistics for Economics and Business,* Second Edition. Izmir (Turkey): Springer.

Maddala, G. S. 1992. *Introduction to Econometrics,* Second Edition. New York: Macmillan Publishing Company.

Pindyck, R. S and D. L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts,* Fourth Edition. New York: McGraw Hill.

Goldberger, A. S. 1998. *Introductory Econometrics***.** Cambridge: Harvard University Press.

Levine, David M., Timothy C. Krehbiei, Mark L. Berenson and P. K. Viswanathan. 2009. *Business Statistics*, Fifth Edition. New Delhi: Pearson Education.

Webster, Allen L. 1998. *Applied Statistics for Business and Economics,* Third Edition. New Delhi: Tata McGraw-Hill.

BLOCK - II

LINEAR REGRESSION

NOTES

UNIT 3 SIMPLE LINEAR REGRESSION

Structure

3.0 INTRODUCTION

In statistics, Simple Linear Regression (SLR) is a linear regression model with a single explanatory variable. , i.e., it concerns two-dimensional sample points with one independent variable and one dependent variable (conventionally, the x and y coordinates in a Cartesian coordinate system) and finds a linear function (a non-vertical straight line) that, as accurately as possible, predicts the dependent variable values as a function of the independent variable. The adjective simply related to the fact that the outcome variable is towards to a single predictor.

In statistics, Ordinary Least Squares (OLS) is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimising the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset and those predicted by the linear function of the independent variable.

The goodness of fit of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarise the discrepancy between observed values and the values expected under the model in question.

NOTES

A statistical hypothesis is a hypothesis that is testable on the basis of observed data modelled as the realised values taken by a collection of random variables. A set of data is modelled as being realised values of a collection of random variables having a joint probability distribution in some set of possible joint distributions. The hypothesis being tested is exactly that set of possible probability distributions. A statistical hypothesis test is a method of statistical inference.

Measurement scale, in statistical analysis, the type of information provided by numbers. Each of the four scales (i.e., nominal, ordinal, interval, and ratio) provides a different type of information. Measurement refers to the assignment of numbers in a meaningful way, and understanding measurement scales is important to interpreting the numbers assigned to people, objects, and events.

In this unit, you will study about the introduction to simple linear regression, estimation of model by method of ordinary least squares, goodness of fit, test for hypothesis, scaling and units of measurement.

3.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the basic concept of simple linear regression
- Comprehend the estimation of model by method of ordinary least squares
- Explain about the goodness of fit
- Analyse the test for hypothesis
- Discuss about the scaling and units of measurement

3.2 INTRODUCTION TO SIMPLE LINEAR REGRESSION

The term 'regression' was first used in 1877 by Sir Francis Galton who made a study which showed that the height of children born to tall parents will tend to move back or 'regress' towards the mean height of the population. He designated the word regression as the name of the process of predicting one variable from the another variable. Then came the term multiple regression to describe the process by which several variables are used to predict another. Thus, when there is a well established relationship between variables, it is possible to make use of this relationship in making estimates and forecasts about the value of one variable (the unknown or the dependent variable) on the basis of the other variable/s (the known or the independent variable/s). For example, if one knows the relationship between tensile strength and hardness of aluminium, one can estimate the tensile strength of aluminium giving its hardness. Such predictions or estimates form the basis of many managerial decisions. For example, a banker could predict deposits on the basis of per capita income in the trading area of bank. A marketing manager may

plan his advertising expenditures on the basis of the expected effect on total sales revenue of a change in the level of advertising expenditure. Similarly, a hospital superintendent could project his need for beds on the basis of total population. Such predictions may be made by using regression analysis. An investigator may employ regression analysis to test his theory having the cause and effect relationship. All this explains that regression analysis is an extremely useful tool specially in problems of business and industry involving predictions. It may, however, be remembered that relationships found by regression are relationships of association but not necessarily of cause and effect. As such one should not infer causality from the relationships one finds by regression. The technique of regression analysis is used to determine the statistical relationship between two (or more) variables and to make prediction of one variable on the basis of the other(s).

Assumptions in Regression Analysis

While making use of the regression technique for making predictions it is always assumed:

(*i*) That there is an actual relationship between the dependent and independent variables;

(*ii*) That the values of the dependent variables are random but the values of the independent variables are fixed quantities without error and are chosen by the experimentor;

(*iii*) That there is clear indication of direction of the relationship. This means that dependent variable is a function of independent variable. (For example, when we say that advertising has an effect on sales, then we are saying that sales has an effect on advertising);

(*iv*) That the conditions (that existed when the relationship between the dependent and independent variable was estimated by the regression) are the same when the regression model is being used. In other words, it simply means that the relationship has not changed since the regression equation was computed;

(*v*) That the analysis be used to predict values within the range (and not for values outside the range) for which it is valid.

Simple Linear Regression Model

In case of simple linear regression analysis, a single variable is used to predict another variable on the assumption of linear relationship (*i.e.*, relationship of the type defined by $Y = a + bX$) between the given variables. The variable to be predicted is called the dependent variable and the variable on which the prediction is based is called the independent variable.

Simple linear regression model (or the Regression Line) is stated as under:

$Y_i = a + bX_i + e_i$

where, Y_i is the dependent variable

X_i is the independent variable

e_i is unpredictable random element (usually called as residual or error term)

(*i*) *a* represents the *Y*-intercept (the intercept specifies the value of the dependent variable when the independent variable has a value of zero. But this term has practical meaning only if a zero value for the independent variable is possible).

(*ii*) '*b*' is a constant indicating the slope of the regression line (slope of the line indicates the amount of change in the value of the dependent variable for a unit change in the independent variable).

If the two constants (viz., *a* and *b*) are known, the accuracy of our prediction of *Y* (denoted by and read as *Y* that depends on the magnitude of the values of e_i. If e_i's in the model tend to have very very large values, then our estimates will not be very good but if their values are relatively small, the predicted values ( ) will tend to be close to the true values (Y_i).

3.3 ESTIMATION OF MODEL BY METHOD OF ORDINARY LEAST SQUARES

Let us look at the following equation:

$$Y = \alpha + \beta X + U \qquad (3.1)$$

Where,

U = Stochastic error term

α, β = Parameters to be estimated

The above equation is called a simple linear regression equation. This is so because there is one dependent variable and one independent variable. In case of multiple regressions, there are at least two independent variables. The equation is estimated using the ordinary least squares (OLS) method of estimation. The OLS method of estimation states that the regression line should be drawn in such a way so as to minimize the error sum of squares. The method of least square is explained as follows:

If we plot the scatter of points on the variable X and Y, the scatter may look as shown in Figure 3.1.

NOTES

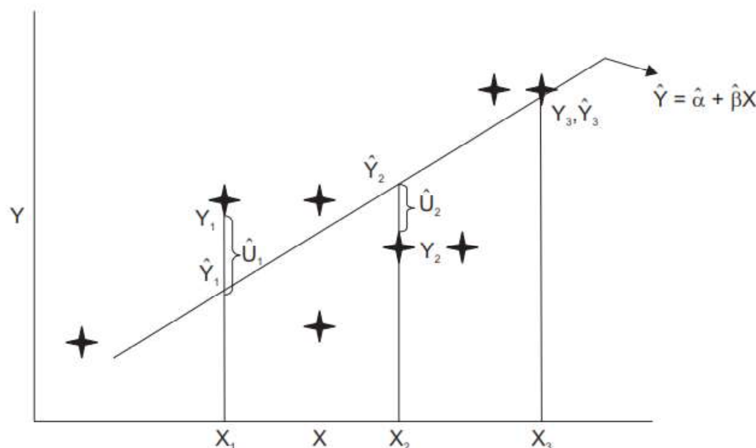


Fig. 3.1 *Method Estimation of Ordinary Least Squares*

Let us assume that $\hat{\alpha}$ and $\hat{\beta}$ are the OLS estimates of α and β respectively. Then, the estimated regression line $\left(\hat{Y} = \hat{\alpha} + \hat{\beta}X\right)$ would look as given in the Figure 3.1. Now corresponding to X_1, there is an observed Y_1 and an estimated value as $\hat{Y}_1$. Therefore, the error is given by $\hat{U}_1 = Y_1 - \hat{Y}_1$ which is positive. Similarly, corresponding to X_2 we have observed Y_2 and estimated $\hat{Y}_2$ and the error is given by $\hat{U}_1 = Y_1 - \hat{Y}_1$ which is negative. Now, for the given value of X_3, the values of Y_3 and $\hat{Y}_3$ are equal as these points lie on the estimated regression line. Therefore, the error is zero. Now the error sum of squares would be given by:

$$\sum_{i=1}^{n} \hat{U}_i^2 = \Sigma (Y - \hat{Y})^2 = \Sigma(Y - \hat{\alpha} - \hat{\beta}X)^2 \qquad (3.2)$$

As mentioned earlier, OLS method aims at minimizing the error sum of square. Therefore, by taking the partial derivative of the above expression with respect to $\hat{\alpha}$ and $\hat{\beta}$ and setting the resulting expression to zero, we get the following:

$$\Sigma Y = n\hat{\alpha} + \hat{\beta}\Sigma X \qquad (3.3)$$

$$\Sigma XY = \hat{\alpha}\Sigma X + \hat{\beta}\Sigma X^2 \qquad (3.4)$$

(We have puposely ignored the derivations and have assumed that the second order conditions for minimization are satisfied.)

The above two equations (3.3 and 3.4) are called normal equations and using algebraic manipulations it can be shown that the OLS estimates of α and β are given as:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \qquad (3.5)$$

$$= \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} \quad (3.6)$$

NOTES

Once $\hat{\beta}$ is estimated, the value of α may be computed as,

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \quad (3.7)$$

After having estimated the regression equation, the estimate of the error (residual) term is obtained as $\hat{U} = Y - \hat{Y}$ where $\hat{U}$ is equal to the estimated value of the error them, Y is the observed value of the dependent variable and $\hat{Y}$ is the estimated value of the dependent variable Y. The estimate of the variance of the error term is given by:

$$V(\hat{U}) = \hat{\sigma}_U^2 = \frac{\sum_{i=1}^{n} \hat{U}_i^2}{n - k} \quad (3.8)$$

Its square root gives the standard error of estimate of the regression equation which is given below:

$$\text{Standard error of estimate} = \hat{\sigma}_U = \sqrt{\frac{\sum_{i=1}^{n} \hat{U}_i^2}{n - k}} \quad (3.9)$$

In the above expression, n and k denote the sample size and the number of parameters to be estimated in a given regression. The standard error of estimates indicates how close the scatter of the points is to the regression line. However, this measure suffers from the defect that it depends upon the units of measurement and, therefore, the fit of the two regression equations with different standard errors of estimates cannot be compared.

Example 3.1: Obtain the usual regression results from following data of 20 pairs of observation *x* on *y*.

$\Sigma x_i = 228$ $\Sigma y = 3121$ $\Sigma x_i y_i = 38977$ $\Sigma x_i^2 = 3204$ $\Sigma x_i y_i = 3347.60$ $\Sigma x_i^2 = 604.8$
$\Sigma y_i^2 = 19837$

We are supposed to fit a linear relation between *y*(Dependent) *x*(Explanatory)

$$\Sigma x_i = 228 \quad n = 20 \quad \bar{x} = 11.4$$
$$\Sigma y_i = 3721 \quad n = 20 \quad \bar{y} = 156.05$$

Estimation of *a* and β

Equation 3.5 $\quad \beta = \frac{\Sigma x_i y_i}{\Sigma x_i^2} \quad \frac{3347.6}{604.8} = 5.54$

$$\hat{a} = \bar{y} - \beta\bar{x}$$
$$a = 156.05 - 5.54(11.4)$$
$$a = 156.05 - 63.156$$

$a = 92.894$

There are estimated regression line is

$$y = a + \beta x \quad \text{(Equation 3.7)}$$

$$y = 92.894 + 5.54x$$

NOTES

Standard Error of estimate

δ_u^2

SE of $\hat{a}$

$$\text{Var } \hat{a} = \delta_u^2 \left(\frac{\sigma x_i^2}{n\Sigma x_i} \right)$$

$$\delta_u^2 = \frac{\Sigma x_i^2}{n-2} = \frac{\Sigma y_i^2 - \beta^2 \Sigma x_i^2}{n-2} = \frac{19837 - (5.54)^2 (604.8)}{20-2} = 70.32$$

$$\text{Var } \hat{a} = \delta_u^2 \left(\frac{\Sigma x_i^2}{n\Sigma x_i^2} \right) \; 70.82 \left(\frac{3204}{20(604.8)} \right) = 18.75$$

$$\text{SE} = \sqrt{\text{Var } \hat{a}} = \sqrt{18.75} = 4.33$$

$$\text{Var } \beta = \frac{\delta_4^2}{\Sigma x_1^2} = \frac{70.82}{604.8} = .117$$

$$\text{SE} = \sqrt{\text{Var} \beta} = \sqrt{.117} = 0.34$$

Example 3.2: Following table give the gross national product (x) and demand for food (y). Estimate the food function and standard eror regression $y = a + \beta_x + u$

n	y_i	x_i	x^2	xy_i	y_i	x_i	$(y-\bar{y})(x_i-\bar{x})$	(x_i^2)	$\hat{y}_i$	e_i	e_i^2
					$y_i - \bar{y}$	$x_i - \bar{x}$	xy_i	$(x_1 - x_2)^2$	$a + \beta x_i$	$y_i - \hat{y}_i$	
1	6	50	2500	300	–2.8	–9.6	26.88	92.16	6.84	–0.85	.72
2	7	52	2704	364	–1.8	–7.6	13.68	57.76	7.25	–0.25	.06
3	8	55	3025	440	–0.8	–4.6	3.68	21.16	7.86	0.13	.02
4	10	59	3481	590	1.2	–0.6	–0.72	0.36	8.67	1.32	1.74
5	8	57	3249	456	–0.8	–2.6	2.08	6.76	8.27	–0.27	0.07
6	9	58	3364	522	0.2	–1.6	–0.32	2.56	8.47	0.53	0.28
7	10	62	3844	620	1.2	2.4	2.88	5.76	9.28	0.71	0.50
8	9	65	4225	585	0.2	5.4	1.08	29.16	9.87	–0.90	0.81
9	11	68	4624	748	2.2	8.4	18.48	70.56	10.51	0.49	0.24
10	10	70	4900	700	1.2	10.4	12.48	108.16	10.91	–0.91	.83
$n = 10$	88	596	35916	5325			80.22	394.4			5.27
$\bar{y}$	$\bar{y} = \frac{88}{10}$	$\bar{x}_1 = \frac{596}{10}$									
	$\bar{y} = 8.8$	59.6									

$\Sigma x_i = 596 \quad \Sigma x_i = 88 \quad \Sigma x_i y_i = 5325 \quad \Sigma x_i^2 = 35916 \quad \Sigma x_i y_i = 80.22 \quad \Sigma x_i^2 = 394.4$

Therefore

$$\beta = \frac{\Sigma x_i y}{\Sigma x_i^2} = \frac{80.22}{394.4} = .2033$$

NOTES

$$\hat{a} = \bar{y} - \beta\bar{x} = \hat{a} = 8.8 - (.2033)(59.6)$$

$$\bar{a} = 8.8 - 12.12$$

$$\bar{a} = -3.316 \text{ or } -3.32$$

$$\hat{y}_i = a + \beta + u$$

$$= -3.32 + .2033\ x_i$$

$$\delta_u^2 \frac{Ec_i^2}{n-2} = \frac{5.27}{10-2} = \frac{5.27}{8} = .6587$$

$$\text{Var } (\hat{a}) = \delta_u \left(\frac{\Sigma x_i^2}{n\Sigma x_i^2}\right).6587\left(\frac{35916}{10(3094.4)}\right)$$

$$.6587\left(\frac{35916}{3944}\right) = .6587(9.106)$$

$$= 5.999$$

$$\text{SE} = \sqrt{\text{Var}\,\hat{a}} = \sqrt{5.999} = 2.449 \text{ or } 2.45$$

$$\text{Var}\,\beta \frac{\delta_u^2}{\Sigma x_i^2} = \frac{.6587}{394.4} = .00167$$

$$\text{SE}\ \sqrt{\text{Var}\,\beta} = \sqrt{.00167} = 0.41$$

$$\Sigma y_i = \Sigma y_i^2 - n\bar{y}^2$$

$$R^2 = \frac{\beta^2(\Sigma x_i^2)}{\Sigma y_i^2} = \frac{.2033^2(394.4)}{21.60} = .756$$

$$R = \sqrt{R^2}$$

Advantages of OLS

We have learnt in the previous section that Ordinary Least Squares (OLS) regression is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable. The method estimates the relationship by minimizing the sum of the squares in the difference between the observed and predicted values of the dependent variable configured as a straight line. Let us have a look at its advantages.

- The solutions work out well for regression model.
- It is fast to compute an ordinary least square estimator.
- When the assumptions are met, it can be more powerful than other regression methods.

- It is familiar to most econometricians.
- The parametric form makes it relatively easy to interpret in comparison to Maximum Likelihood estimator.
- The OLS estimator minimizes the average squared difference between the actual values of Yi and the prediction ('predicted value') based on the estimated line.

3.3.1 Statistical Properties of OLS

Let's have a look at the statistical properties of OLS.

Property 1: Linear

Linear property is more concerned with the estimator rather than the original equation that is being estimated. It is assumed the regression analysis focus that the linear regression should be 'linear in parameters.' However, the *linear* property of OLS estimator means that OLS belongs to that class of estimators, which are linear in Y, the dependent variable. Note that OLS estimators are linear only with respect to the dependent variable and not necessarily with respect to the independent variables.

Property 2: Un-biasedness

In a regression equation, an error term is associated with the regression equation that is estimated. This makes the dependent variable also random. If an estimator uses the dependent variable, then that estimator would also be a random number. Therefore, before describing what un-biasedness is, it is important to mention that un-biasedness property is a property of the estimator and not of any sample.

Un-biasedness is one of the most desirable properties of any estimator. The estimator should ideally be an unbiased estimator of true parameter/population values.

Consider a simple example: Suppose there is a population of size 1000, and a sample of 50 is taken from this population to estimate the population parameters. Every time a sample is taken, it will have different set of 50 observations and, hence, it estimates different values of β_o *and* β_1.

The unbiasedness property of OLS method says that when a sample of 50 is taken repeatedly, then after some repeated attempts, it is found that the average of all the β_o *and* β_1 from the samples will be equal to the actual (or the population) values of β_o *and* β_1.

Mathematically,

$E(b_o) = \beta_o$

$E(b_i) = \beta_i$

Here, 'E' is the expectation operator.

The un-biasedness property of OLS in Econometrics is the basic minimum requirement to be satisfied by any estimator. However, it is not sufficient most of the times in real-life applications as we may not find similar repeated samples.

Property 3: Best Minimum Variance

The efficient property of any estimator says that the estimator is the *minimum variance unbiased* estimator. Therefore, if all the unbiased estimators are taken of the unknown population parameter, the estimator will have the least variance. The estimator that has less variance will have individual data points closer to the mean. As a result, they will be more likely to give better and accurate results than other estimators having higher variance. In short:

1. If the estimator is unbiased but doesn't have the least variance – it's not the best!
2. If the estimator has the least variance but is biased – it's again not the best!
3. If the estimator is both unbiased and has the least variance – it's the best estimator.

Now, talking about OLS, OLS estimators have the *least variance* among the class of all *linear unbiased* estimators. So, this property of OLS regression is less strict than efficiency property. Efficiency property says least variance among all unbiased estimators, and OLS estimators have the least variance among all linear and unbiased estimators.

Property 4: Asymptotic Un-biasedness

This property of OLS says that as the sample size increases, the biasedness of OLS estimators disappears.

Property 5: Consistency

An estimator is said to be consistent if its value approaches the actual, true parameter (population) value as the sample size increases. An estimator is consistent if it satisfies two conditions:

a. It is asymptotically unbiased
b. Its variance converges to 0 as the sample size increases.

Both these hold true for OLS estimators and, hence, they are consistent estimators. For an estimator to be useful, consistency is the minimum basic requirement. Since there may be estimators, where, asymptotic efficiency also is considered. Asymptotic efficiency is the sufficient condition that makes OLS estimators the best estimators.

3.3.2 Numerical Properties of OLS

Now that we have seen the statistical properties, let us study the numerical properties of OLS.

1. The OLS estimators are expressed solely in terms of the observables (i.e.samples)quantities, hence they can be easily computed.
2. OLS estimators are point estimators which means that in a sample each estimator will provide only a single value for a corresponding population parameter.

NOTES

3. Once we have OLS estimator from sample data the sample regression lines can be easily obtained. The regression line will be having following numerical properties:
   (a) The regression line passes through the sample mean of Y and X
   (b) The mean value of the estimated Y= y is equal to the mean value of the actual Y
   (c) The mean value of the residuals u_i is zero
   (d) The residuals u_i are uncorrelated with the predicted Y_i.
   (e) The residuals u_i are uncorrelated with Xi that is $\cong u_i X_i = 0$

Check Your Progress

1. What is regression?
2. Elaborate on the simple linear regression.
3. What is dependent and independent variable and why is called so?
4. Explain about the linear property in the case of OLS.

3.4 GOODNESS OF FIT

Tests of goodness of fit enable us to determine how good a fit is between the observed frequencies and the corresponding expected frequencies. Whereas the former are the outcomes based on observing a sample of size *n,* the latter are obtained from the hypothesized population from which the sample is drawn.

It may be appreciated that a given population can always be reasonably expected to obey a certain probability distribution. We may thus be interested in verifying on the basis of sample information whether it can be adequately explained by a specified theoretical probability distribution.

For example, majority of the tests developed are based on the assumption of the population being normal. If this assumption is in doubt, an appropriate testing procedure is needed to verify that a given population is actually described by the normal probability model. The testing procedure used for such verification is called the test of goodness of fit. Thus, the basic function of this test is to determine whether a population possesses a specified theoretical probability distribution.

Setting the test of goodness of fit requires taking the following two steps:

(i) In the first place, the test requires formulating the null hypothesis stating that a given population obeys a specific probability distribution (such as the uniform, binomial, Poisson, or normal). A random sample of size *n* is then selected from the population and the observed frequencies falling in various classes are recorded. The resultant distribution is known as the *empirical frequency distribution* for the sample.

(ii) In the second place, the theoretical distribution underlying the null hypothesis is identified and the probability values for the various classes

found. These probabilities, when multiplied by the sample size, yield the corresponding expected class frequencies, which constitute our *theoretical frequency distribution* for the sample.

NOTES

In the end, the test of goodness of fit compares the empirical frequency distribution with the theoretical frequency distribution by computing the X^2 value. The decision to accept or reject H_0 is taken by applying the decision rule discussed in the previous section in exactly the same way as demonstrated here.

Let us see how the tests of goodness of fit are applied to various population distributions, such as, binomial, Poisson, and normal. It will be discovered, as we proceed, that these tests enable us to verify, on the basis of a sample drawn from it, whether a given population follows a certain distribution.

Testing the Goodness of Fit of a Binomial Distribution

Let 5 coins be tossed 100 times. The number of heads coming up in each trial is noted. When tabulated, the results observed are found distributed as in Col. (2) of Table 3.1, which are the observed frequencies O_i's. Since it is a case of binomial experiment, it interests us to know if the distribution of the number of heads in n = 100 trials follows the binomial distribution. Thus, the null hypothesis H_0 is that the sample distribution agrees with the hypothesized (binomial) distribution.

Table 3.1 *Computation of Expected Frequencies Using the Binomial Rule*

No. of Heads	O_i		$b(X; 5, 1/2)$	$e_i = 100\, b(X; 5, 1/2)$		$(O_i - e_i)^2/e_i$
(1)	(2)		(3)	(4)		(5)
0	2		0.0312	3.12		
		8			18.74	6.155
1	6		0.1562	15.62		
2	25		0.3125	31.25		1.25
3	40		0.3125	31.25		1.922
4	20		0.1562	15.62		
		27			18.74	3.641
5	7		0.0312	3.12		
						$\chi^2 = 12.97$

To determine the expected frequencies, a binomial distribution is fitted on the sample data. The assumption being that all the five coins are fair, the probability p of

getting a head in the toss of a single coin is $p = 1/2$. Therefore, the binomial distribution to be used is

$$b\left(X; 5, \frac{1}{2}\right) = \frac{5!}{X!\,(5-X)!}\left(\frac{1}{2}\right)^{X}\left(\frac{1}{2}\right)^{5-X},$$

NOTES

in which X is the binomial random variable representing the number of heads.

When solved for different values of X from 0 to 5, we obtain the binomial probabilities for each value of X as in Col. (3). These probabilities when multiplied by $n = 100$ give the expected frequencies e_i as recorded in Col. (4). The X^2 value is then computed, as in Col. (5), to take decision on H_0 at ± = 0.05 level of significance.

It may be noted that two classes are lost in regrouping because e_i for the first and the last class are less than 5. Since $k = (6 - 2) = 4$, the *df* $v = k - 1 = 4 - 1 = 3$. At $\alpha = 0.05$ and for $v = 3$, $\chi 2_{0.05}$=7.81. The computed X^2 value $\chi^2 = 12.97$ being greater than H_0 is rejected at 0.05 level of significance. It means the sample distribution is not in agreement with the binomial distribution.

Testing the Goodness of Fit of a Poisson Distribution

Consider for example a publishing company, Vikas Publishing House, who have got a 500 page book composed for printing. Before the final printing, the draft manuscript is sent for proof reading. The proof reader discovers varying number of misprints X per page. When tabulated, these are found distributed as in Cols. (1) and (2), respectively, of Table 3.2.

Since the problem involves a Poisson experiment, our interest may be in knowing if the number of mistakes per page follows the Poisson distribution. As before, the null hypothesis H_0 is that the sample distribution agrees with the hypothesized Poisson distribution.

Table 3.2 *Computation of Expected Frequency Using the Poisson Distribution*

Misprints per Page, X	O_i		X_iO_i	$P(X; 0.9)$	$e_i = (500)\, P(X; 0.9)$		$(O_i - e_i)^2/e_i$
(1)	(2)		(3)	(4)	(5)		(6)
0	221		0	0.406	203		1.60
1	167		167	0.366	183		1.40
2	70		140	0.164	82		1.76
3	30		90	0.050	25		1.00
4	7		28	0.011	5		
		12				7	3.57
5	5		25	0.003	2		
			450				$\chi^2 = 9.33$

Testing H_0 requires fitting a Poisson distribution to the sample data by obtaining the corresponding e_i's. The probabilities for different values of X can then be found by using the Poisson distribution

NOTES

$$P(X;\mu) = \frac{e^{-\mu}\mu^X}{X!},$$

in which $X = 1,2,3,....$

Since the average number of misprints μ is not specified, it may be estimated from the sample observed frequencies as

$$\mu = \frac{\sum X_i O_i}{\sum O_i},$$

which, according to Col. (3), comes to μ = 450/500 = 0.9. For μ = 0.9, the probabilities for different values of X obtained by using a cumulative poisson probabilities table are as in Col. (4), and the corresponding e_i's as in Col. (5), of Table 3.2.

Since one class is lost in grouping, the number of classes $k = 5$ and $m = 1$, so that $v = k - m - 1 = 3$. For ± = 0.05 and

$v = 3$, Since the computed X^2 value $\chi^2 = 9.33$ is more than H_0 is rejected. It may thus be concluded that the sample distribution of misprints per page is not in agreement with the Poisson distribution.

Testing the Goodness of Fit of a Normal Distribution

Consider the hypothetical frequency distribution of a sample of 200 workers employed in spinning mills, as given in Table 3.3. These data can be used to test the hypothesis that the population to which the sample belongs, follows a normal distribution. Thus, the null hypothesis H_0 is that the sample distribution is in agreement with the normal distribution. Testing H_0 requires fitting a normal distribution to the sample data so as to determine the corresponding e_i's.

Since the population mean μ and the standard deviation σ are not specified, these are estimated by the corresponding sample values as the point estimates. With sample mean and sample standard deviation $s = 4.18$, the probabili-ties for each class are obtained by using the normal area table in the manner described below. When the class probabilities are multiplied by the sample size n, the result-ant frequencies are the required e_i's.

Determining probabilities for any class requires computation of Z values corresponding to the boundaries of that class. *For example*, the Z values for the second class are

$$z_1 = \frac{24 - 33.4}{4.18} = -2.249 \quad \text{and} \quad z_2 = \frac{28 - 33.4}{4.18} = -1.292.$$

NOTES

Table 3.3 *Distribution of 200 Workers According to their Daily Earnings*

Daily Earnings (₹), X	No. of Workers, O_i
(1)	(2)
20–24	8
24–28	15
28–32	40
32–36	90
36–40	35
40–44	9
44–48	3
	200

Then, the desired probability for the second class is

$$P(24 < X < 28) = P(-2.249 < Z < -1.292)$$
$$= P(Z < -1.29) - P(Z < -2.25)$$
$$= 0.0985 - 0.0122 = 0.0863,$$

and the expected frequency $e_i = 200\ (0.0863) = 17.26$. We proceed the same way in the case of all classes, except for the first and last class.

Using a similar method for obtaining the probabilities for the first and the last class ignores the extreme tail-end areas under the normal curve. To take care of the probabilities allocated by the two tail-ends of the normal curve, it is necessary to take the first class as ranging from–” to 24 and the last class as ranging form 44 to +”. Accordingly, the probability of the first class is $P(X< 24)$ and that of the last class is $P(X > 44)$. Computing the Z values for $X = 24$ and $X = 44$, the two probabilities are then obtained as usual.

Thus, the probabilities for each class given in Col. (3) of Table 3.4, the expected frequencies e_i are obtained as in Col (4). With corresponding observed frequencies O_i given in Col. (2), the Ç^2 value is computed as in Col. (5).

Table 3.4 *Computation of Expected Frequencies Using the Normal Distribution*

Daily Earnings (₹)	No. of Workers (O_i)		$n(X;\ 33.4, 4.18)$	$e_i = 200 \times n(X;\ 33.4, 4.18)$		$(O_i - e_i)^2/e_i$
(1)	(2)		(3)	(4)		(5)
20–24	8		0.0116	2.32		
		23			19.70	0.60
24–28	15		0.0863	17.26		4.30
28–32	40		0.2722	55.44		4.31
32–36	90		0.3617	72.34		1.20
36–40	35		0.2106	42.12		
40–44	9		0.0515	10.30		
		12			11.40	0.03
44–48	3		0.0055	1.10		
	200			200		$\chi^2 = 10.44$

NOTES

Two classes, first and the last, are lost in grouping so that now $k = 5$. With $m = 2$, $v = k - m - 1 = 2$. For $v = 2$ and $\alpha = 0.05$, the tabulated X^2 value is Since the computed X^2 value $\chi^2 = 10.44$ is much higher than H_0 is rejected. It means that the sample data do not come from a normal population. In other words, the sample information does not provide adequate evidence in support of the population being normal.

3.5 TEST FOR HYPOTHESES

Let us look at hypothesis tests on the regression coefficients of simple linear regression. It is possible to carry out tests where one can assume that the random error term, $\in$, is distributed normally and independently, having mean as zero and variance of σ^2. Let us look at these tests in detail.

We employ *t* tests for testing hypothesis for the regression coefficients got from simple linear regression. The statistic which is based on the *t*-distribution gets employed for testing the two-sided hypothesis that the true slope, β_1, is equal to a constant value, $\beta_{1,0}$. For the hypothesis test the statement will be:

$$H_0 : \beta_1 = \beta_{1,0}$$
$$H_1 : \beta_1 \neq \beta_{1,0}$$

This test's test statistic will be:

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se\left(\hat{\beta}_1\right)}$$

In the above:

$\hat{\beta}_1$ depicts the least square estimate of $\hat{\beta}_1$ with $se\left(\hat{\beta}_1\right)$ depicting its standard error.

Here is how to calculate the value of $se\left(\hat{\beta}_1\right)$:

$$se\left(\hat{\beta}_1\right) = \sqrt{\frac{\dfrac{\sum_{i=1}^{n} e_i^2}{n-2}}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$$

The test statistic, T_0, follows a $(n-2)$ distribution with degrees of freedom. In it, n depicts the total number of observations. There will be an accepting of the null hypothesis, H_0, when the test statistic's calculated value is such that:

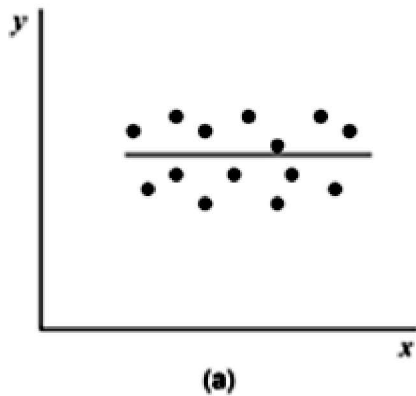
$$-t_{\alpha/2,n-2} < T_0 < t_{\alpha/2,n-2}$$

NOTES

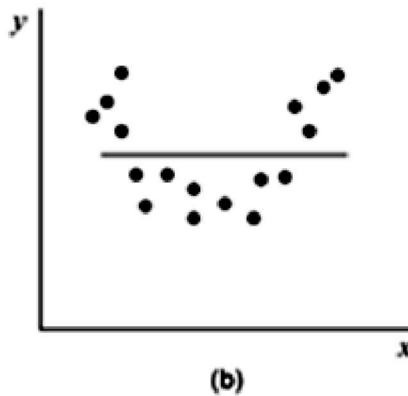
In which:

$t_{\alpha/2,\,n-2}$ and $-t_{\alpha/2,\,n-2}$ are the two-sided hypothesis' critical values. Also, $t_{\alpha/2,\,n-2}$ represents the percentile of the t distribution which corresponds to a cumulative probability of $(1-\alpha/2)$ with α being the significance level.

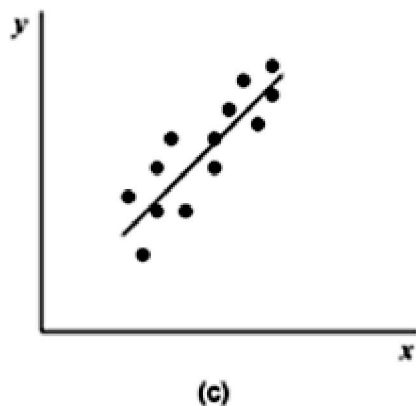
If zero is the used value of $\beta_{1,0}$, the hypothesis will test for the significance of regression. That is, the test shows whether the fitted regression model is of value in explaining variations in the observations or whether a regression model is being imposed in a case where there is no true relationship between x and Y. In case $H_0 : \beta_1 = 1$ cannot be rejected, it means that no linear relationship exists between x and Y. It is possible to get such a result in case where the scatter plots are given (a) and (b) as given below. The figure labelled (a) is representative of a case in which there is no model in existence for the data that has been observed and so if one attempts to fit a regression model here, it will actually be associated with random variation or noise. The figure labelled (b) depicts a case in which the true relationship between x and Y is not linear. The figures (c) and (d) illustrate the case in which $H_0 : \beta_1 = 1$ has been rejected, showing that there does not exist any model between x and Y. Figure (c) displays a case in which the linear model is sufficient. The figure labelled (d) illustrates a case in which there might be need for a higher order model.



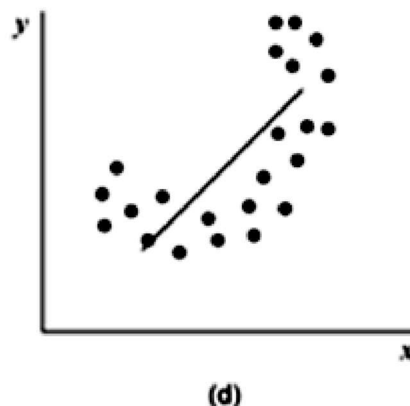
(a)



(b)



(c)



(d)

NOTES

It is possible to employ a similar procedure for the purpose of testing the hypothesis on the intercept. In such a case, the test statistic employed would be:

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{se\left(\hat{\beta}_0\right)}$$

In the above:

$\hat{\beta}_0$ is the least square estimate of β_0, and $se\left(\hat{\beta}_0\right)$ is its standard error that is calculated with:

$$se\left(\hat{\beta}_0\right) = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right]}$$

3.6 SCALING AND UNITS OF MEASUREMENT

In general, all data, whether qualitative or quantitative, is measured in some form. Even the discrete quantitative data which is counted can fit into some form of measurement. There are four widely accepted levels of measurement. These levels, from the weakest on the one extreme to the strongest on the other, in order are: Nominal scale, Ordinal scale, Interval scale and Ratio scale. Before discussing these various measurement levels, let us look at some of the attributes possessed by these scales.

(a) **Magnitude:** It is the quantitative value that exists or is assigned to an attribute or characteristic and such values, when compared, will determine whether the value of a given attribute in one case is greater than, equal to or less than the value of the same attribute in another case. For instance, if student *X* gets 100 per cent marks in the final examination in a course and student *Y* gets 40 per cnet in the same exam, then student *X* may be considered as more knowledgeable in that area than student *Y*.

(b) **Equal intervals:** Some measurement scales are constructed in such a manner that the magnitude of an interval between any two points along the scale has the same value or the same magnitude within the same interval of any other two points along the same scale. For instance, the difference in heights of students between 60 inches and 63 inches is the same in magnitude as the difference between 70 inches and 73 inches. This means that the value of the magnitude is 3 inches, no matter where such interval is measured

on the scale. There may be some exceptions to this rule. For instance, the value of the difference between the IQ of 180 and 190 may be different than the value of the difference of an IQ of 80 and 90, even though, numerically both these differences have the same value.

NOTES

(c) **Absolute zero point:** The third attribute of the measurement scale is the presence or absence of the zero point where the attribute has no value at all. For instance, the characteristic of height of a person does not have an absolute zero point, since positive quantitative value of the attribute always exists, no matter what the age of the person may be. On the other hand, the number of TV sets in a family can have an absolute zero value if the family has no TV set at all. In some unique cases we may assign a zero value to an attribute for qualitative comparison purposes even when the value of such an attribute is a positive quantitative number. For instance, we may say that an unintelligent person has zero intelligence, even though it does not mean absolute zero.

In light of these three attributes, let us now consider and discuss the four measurement scales.

(i) Nominal scale: Applied to qualitative data only, it is also known as classificatory scale, where the objects or items are classified into various discrete and distinct groups or categories without any ranking or order associated with such classified data. It does not possess any of the three attributes discussed earlier: magnitude, equal intervals and absolute zero point. It is the weakest form of measurement so that some statisticians do not consider it as a scale at all. Examples of nominal scale would be categorizing people according to their religion such as the Christian, Muslim, Hindu and so on, or according to their political affiliation such as the Democrat, Republican or Socialist. Other categories of nominal scale may be smoking versus non-smoking, ownership of the house versus no ownership of the house, and so on.

(ii) Ordinal scale: Also known as ranking scale, it possesses only the attribute of magnitude. This means that various categories of items can be compared with each other only in order of rank assigned to these categories. However, these ranks only indicate as to which category is greater or better, but does not indicate the magnitude of the difference among these categories. For instance, the students in a class may be categorized according to their grades of *A*, *B*, *C*, *D* and *F* where *A* is better than *B*, and so on, and the classification is from the highest grade to the lowest grade. Another example of ordinal scaling would be the classification of teaching faculty ranks in the colleges as full professors, associate professors, assistant professors and instructors.

NOTES

(iii) Interval scale: The interval scale measures the values of quantitative random variables and identifies not only as to which category is greater or better but also by how much. It is a stronger form of measurement and possesses two of the attributes which are magnitude and equal intervals. It does not possess, however, the absolute zero point. Measurements of height, weight and time are all examples of interval scale.

(iv) Ratio scale: The ratio scale is also used for measurement of quantitative random variables, but it differs from interval scale in that it has a true zero point, meaning that the values of such variables can be zero. It makes the mathematical manipulations easier such as divisions and multiplications. Examples of ratio scale are physical measurements including temperature, number of students registered in various classes, and so on. The temperature can be zero which means the total absence of heat and it is also possible that zero students are registered for a given class. Similarly, heights and weights, though considered in interval scale, can have hypothetical zero values.

These measurement scales assist in designing survey methods for the purpose of collecting relevant data.

Check Your Progress

5. State the basic function of the goodness of fit test.
6. What is test of hypotheses depend upon the regression analysis?
7. Give the names four levels of measurement.
8. Explain about the ordinal scale.
9. What is interval scale?

3.7 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The term 'regression' was first used in 1877 by Sir Francis Galton who made a study which showed that the height of children born to tall parents will tend to move back or 'regress' towards the mean height of the population. He designated the word regression as the name of the process of predicting one variable from the another variable. Then came the term multiple regression to describe the process by which several variables are used to predict another. Thus, when there is a well established relationship between variables, it is possible to make use of this relationship in making

estimates and forecasts about the value of one variable (the unknown or the dependent variable) on the basis of the other variable/s (the known or the independent variable/s).

2. In case of simple linear regression analysis, a single variable is used to predict another variable on the assumption of linear relationship (*i.e.*, relationship of the type defined by $Y = a + bX$) between the given variables.
3. The variable to be predicted is called the dependent variable and the variable on which the prediction is based is called the independent variable.
4. Linear property is more concerned with the estimator rather than the original equation that is being estimated. It is assumed the regression analysis focus that the linear regression should be 'linear in parameters.' However, the *linear* property of OLS estimator means that OLS belongs to that class of estimators, which are linear in Y, the dependent variable.
5. Tests of goodness of fit enable us to determine how good a fit is between the observed frequencies and the corresponding expected frequencies. Whereas the former are the outcomes based on observing a sample of size *n,* the latter are obtained from the hypothesized population from which the sample is drawn.
6. Hypothesis tests on the regression coefficients of simple linear regression. It is possible to carry out tests where one can assume that the random error term, $\in$, is distributed normally and independently, having mean as zero and variance of σ^2.
7. There are four widely accepted levels of measurement. These levels, from the weakest on the one extreme to the strongest on the other, in order are: Nominal scale, Ordinal scale, Interval scale and Ratio scale.
8. Ordinal scale: Also known as ranking scale, it possesses only the attribute of magnitude. This means that various categories of items can be compared with each other only in order of rank assigned to these categories.
9. Interval scale: The interval scale measures the values of quantitative random variables and identifies not only as to which category is greater or better but also by how much. It is a stronger form of measurement and possesses two of the attributes which are magnitude and equal intervals.

3.8 SUMMARY

- A hospital superintendent could project his need for beds on the basis of total population. Such predictions may be made by using regression analysis. An investigator may employ regression analysis to test his theory having the cause and effect relationship.

- The technique of regression analysis is used to determine the statistical relationship between two (or more) variables and to make prediction of one variable on the basis of the other(s).
- That the values of the dependent variables are random but the values of the independent variables are fixed quantities without error and are chosen by the experimentor;
- In case of simple linear regression analysis, a single variable is used to predict another variable on the assumption of linear relationship (*i.e.*, relationship of the type defined by $Y = a + bX$) between the given variables.
- The variable to be predicted is called the dependent variable and the variable on which the prediction is based is called the independent variable.
- In case of multiple regressions, there are at least two independent variables. The equation is estimated using the ordinary least squares (OLS) method of estimation. The OLS method of estimation states that the regression line should be drawn in such a way so as to minimize the error sum of squares.
- Ordinary Least Squares (OLS) regression is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable. The method estimates the relationship by minimizing the sum of the squares in the difference between the observed and predicted values of the dependent variable configured as a straight line.
- The OLS estimator minimizes the average squared difference between the actual values of Yi and the prediction ('predicted value') based on the estimated line.
- In a regression equation, an error term is associated with the regression equation that is estimated. This makes the dependent variable also random. If an estimator uses the dependent variable, then that estimator would also be a random number.
- The un-biasedness property of OLS in Econometrics is the basic minimum requirement to be satisfied by any estimator. However, it is not sufficient most of the times in real-life applications as we may not find similar repeated samples.
- The efficient property of any estimator says that the estimator is the *minimum variance unbiased* estimator. Therefore, if all the unbiased estimators are taken of the unknown population parameter, the estimator will have the least variance.
- The estimator that has less variance will have individual data points closer to the mean. As a result, they will be more likely to give better and accurate results than other estimators having higher variance.

NOTES

- An estimator is said to be consistent if its value approaches the actual, true parameter (population) value as the sample size increases.
- OLS estimators are point estimators which means that in a sample each estimator will provide only a single value for a corresponding population parameter.
- Tests of goodness of fit enable us to determine how good a fit is between the observed frequencies and the corresponding expected frequencies. Whereas the former are the outcomes based on observing a sample of size *n,* the latter are obtained from the hypothesized population from which the sample is drawn.
- The test of goodness of fit compares the empirical frequency distribution with the theoretical frequency distribution by computing the X^2 value. The decision to accept or reject H_0 is taken by applying the decision rule discussed in the previous section in exactly the same way as demonstrated here.
- Hypothesis tests on the regression coefficients of simple linear regression. It is possible to carry out tests where one can assume that the random error term, $\in$, is distributed normally and independently, having mean as zero and variance of σ^2.
- If zero is the used value of $\beta_{1,0}$, the hypothesis will test for the significance of regression. That is, the test shows whether the fitted regression model is of value in explaining variations in the observations or whether a regression model is being imposed in a case where there is no true relationship between x and Y.
- There are four widely accepted levels of measurement. These levels, from the weakest on the one extreme to the strongest on the other, in order are: Nominal scale, Ordinal scale, Interval scale and Ratio scale.
- Absolute zero point: The third attribute of the measurement scale is the presence or absence of the zero point where the attribute has no value at all.
- Applied to qualitative data only, it is also known as classificatory scale, where the objects or items are classified into various discrete and distinct groups or categories without any ranking or order associated with such classified data.
- Ordinal scale: Also known as ranking scale, it possesses only the attribute of magnitude. This means that various categories of items can be compared with each other only in order of rank assigned to these categories.
- Interval scale: The interval scale measures the values of quantitative random variables and identifies not only as to which category is greater or better but also by how much. It is a stronger form of measurement and possesses two of the attributes which are magnitude and equal intervals.

3.9 KEY WORDS

NOTES

- **Tests of goodness fit:** It enables us to determine how good a fit is between the observed frequencies and the corresponding expected frequencies.
- **Testing of hypotheses:** Hypothesis tests on the regression coefficients of simple linear regression. It is possible to carry out tests where one can assume that the random error term, ε, is distributed normally and independently, having mean as zero and variance of σ^2.
- **Magnitude:** It is the quantitative value that exists or is assigned to an attribute or characteristic and such values, when compared, will determine whether the value of a given attribute in one case is greater than, equal to or less than the value of the same attribute in another case.
- **Absolute zero point:** Absolute zero point is the third attribute of the measurement scale is the presence or absence of the zero point where the attribute has no value at all.
- **Nominal scale:** Nominal scale is the applied to qualitative data only, it is also known as classificatory scale, where the objects or items are classified into various discrete and distinct groups or categories without any ranking or order associated with such classified data.

3.10 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Define the term regression.
2. Give the assumptions in regression analysis.
3. What is simple linear regression model?
4. Elaborate on the ordinary least squares.
5. Give the advantages of OLS.
6. Explain about the numerical properties of OLS.
7. What do you understand by goodness of fit?
8. Interpret the testing of goodness fit of binomial distribution.
9. Explain the testing the goodness of fit of a poisson distribution.
10. Elaborate on the test for hypotheses.
11. What is equal interval?
12. Explain about the absolute zero point.
13. Define the term nominal scale.

14. Interpret the ratio scale.

Long-Answer Questions

NOTES

1. Briefly explain about the simple linear regression giving assumption with the appropriate examples.
2. Describe the estimation of model by method of ordinary least squares giving advantages.
3. Explain in detail about the statistical and numerical properties of OLS with the help of examples.
4. What is goodness of fit? Explain with the various examples.
5. Analyse the testing the goodness of fit of normal, binomial and poisson distribution.
6. Explain in detail about the test for hypotheses with its characteristics.
7. Discuss in detail about the measurement levels with appropriate examples.
8. Describe the various type of measurement scales.

3.11 FURTHER READINGS

Johnston, J. and John DiNARDO. 1997. *Econometric Methods*, Fourth Edition. New Delhi: Tata McGraw-Hill.

Koutsoyiannis, A.1977. *Theory of Econometrics,* Second Edition. London: The Macmillan Press Ltd.

Özdemir, Durmuº. 2016. *Applied Statistics for Economics and Business,* Second Edition. Izmir (Turkey): Springer.

Maddala, G. S. 1992. *Introduction to Econometrics,* Second Edition. New York: Macmillan Publishing Company.

Pindyck, R. S and D. L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts,* Fourth Edition. New York: McGraw Hill.

Goldberger, A. S. 1998. *Introductory Econometrics***.** Cambridge: Harvard University Press.

Levine, David M., Timothy C. Krehbiei, Mark L. Berenson and P. K. Viswanathan. 2009. *Business Statistics*, Fifth Edition. New Delhi: Pearson Education.

Webster, Allen L. 1998. *Applied Statistics for Business and Economics,* Third Edition. New Delhi: Tata McGraw-Hill.

UNIT 4 MULTIPLE LINER REGRESSION MODEL

NOTES

Structure

4.0 INTRODUCTION

A multiple linear regression model is a linear model that describes how a *y*-variable relates to two or more *x*-variables (or transformations of *x*-variables). In this unit, the model has been explained with the help of several useful equations. The estimators that we create through linear regression give us a relationship between the variables. However, performing a regression does not automatically give us a reliable relationship between the variables. In order to create reliable relationships, we must know the properties of the estimators and show that some basic assumptions about the data are true. Estimators help set the rules of estimation for a set of data. One of the most simplistic estimator used in Ordinary Least Squares (OLS) method. In this unit, we will take up the estimation of Population Regression Function (PRF) through the OLS method. We will study advantages, numerical and statistical properties of OLS along with the assumptions of the classical linear regression model.

Parameter estimates (also called coefficients) are the transformation form of the response associated with a one-unit change of the predictor, all other predictors being held constant. The unknown model parameters are estimated using least-squares estimation.

A goodness-of-fit test, in general, related to the measuring how well do the observed data correspond to the fitted (assumed) model. Like in a linear regression,

in essence, the goodness-of-fit test compares the observed values to the expected (fitted or predicted) values.

R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model. The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model actually increases the model fit.

In this unit, you will study about the introduction of multiple linear regression model, estimation of parameters, properties of OLS estimators, goodness of fit, R^2 and adjusted R^2.

4.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain about the multiple regression model
- Discuss about the estimation of parameter
- Interpret the properties of OLS estimators
- Understand the goodness of fit
- Analyse the R^2 and adjusted R^2

4.2 INTRODUCTION TO MULTIPLE LINER REGRESSION MODEL

In the multiple regression model, there is extension of the simple (two-variable) regression model to further take into account possibility of additional explanatory factors which can systematically affect the dependent variable. The most simple way in which they can be extended is by using the model which is referred to as the three-variable model which uses an additional secondary variable as shown below.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \quad (4.1)$$

In the above, every one of the slope coefficients are partial derivatives of *y* with respect to the variable *x,* that is multiplied by them. So, if x_2 is fixed:

$$\beta_1 = \partial y / \partial x_1$$

The extension enables the consideration of nonlinear relationships, like polynomial in *z* where

$x_1 = z$ *and* $x_2 = z^2$. Here regression will be linear in x_1 and x_2 and nonlinear in *z*.

$$\partial y / \partial z = \beta_1 + 2\beta_2 z$$

NOTES

In this model, the prime assumption involves the independence of the error process *u* and both regressors/explanatory variables:

$$E\,(u \mid x_1, x_2) = 0. \tag{4.2}$$

Assuming a zero conditional mean for the error process is indicative of it not systematically varying with $x's$ or any linear combination of $x's$.

In the statistical way, *u* is independent of the distributions of $x's$.

Let us look at this model in context of k regressors. It will be as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u \tag{4.3}$$

with interpretation of the β coefficients being the same, with every one being a partial derivative of *y* with respect to *x*; holding all other $x's$ constant (ceteris paribus). Here the *u* term is *y*'s nonsystematic which is not linearly related with any $x's$. The dependent variable *y* is considered linearly related with $x's$ and can have any relation with each other till no exact linear dependencies exist amongst the regressors.

So, the independence assumption will be:

$$E\,(u \mid x_1, x_2, ..., x_k) = 0. \tag{4.4}$$

OLS: Mechanics and Interpretation

Let us begin with the three-variable model which was referred earlier at Equation (4.1).

The OLS equation that is estimated has the parameters of interest:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 \tag{4.5}$$

Here, one can define the ordinary least squares criterion in terms of the OLS residuals, calculated by using the expression given below:

$$\min S = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2})^2 \tag{4.6}$$

This expression's minimization is done with respect to each of the three parameters, $\{b_0, b_1, b_2\}$. When there are *k* regressors, these expressions will contain terms in b*k* with the minimization being done with respect to (*k*+1) parameters $\{b_0, b_1, b_2, b_k\}$ so that this becomes feasible, the sample has to be larger than the number of parameters that have to be estimated from the sample ($n > (k + 1)$).

We carry out the minimization through differentiating the scalar S for all $b's$ turn wise, and making the first order condition that results a zero. So we get, (*k*+1) simultaneous equations in (*k*+1) unknowns, called the least squares normal equations.

NOTES

In the three-variable regression model, the normal equations can be written as follows:

$$\begin{aligned}\sum y &= nb_0 + b_1\sum x_1 + b_2\sum x_2 \qquad (\\ \sum x_1 y &= b_0\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1x_2 \\ \sum x_2 y &= b_0\sum x_2 + b_1\sum x_1x_2 + b_2\sum x_2^2\end{aligned} \qquad (4.7)$$

It is possible to interpret the first normal equation as specifying that the regression surface (in 3-space), goes through the multivariate point of means $\{\bar{x}_1, \bar{x}_2, \bar{y}\}$..

It is possible to solve the above three equations uniquely through using normal algebraic techniques/linear algebra, for the estimated least squares parameters.

This will also hold for *k* regressors and (*k*+1) regression parameters.

Fitted values, residuals, and their properties

One has to calculate fitted values/predicted values, post making an estimation of a multiple regression.

In the case of observation *I*, the fitted value will be:

$$\hat{y}_i = b_0 + b_1x_{i1} + b_2x_{i2} + ... + b_kx_{ik} \qquad (4.8)$$

with the residual being the difference of *y*'s actual value and the fitted value as given below:

$$e_i = y_i - \hat{y}_i \qquad (4.9)$$

Sum of the residuals is zero. By construction, they possess zero covariance with each of the x variables, and therefore zero covariance with $\hat{y}$. With the average residual being zero, the regression surface will pass through the multivariate point of means:

$$\{\bar{x}_1, \bar{x}_2, ..., \bar{x}_k, \bar{y}\}.$$

Two instances exist in which simple regression of y on x_1 provides the same coefficient as multiple regression of y on x_1 and on x_2, with respect to x_1.

Generally, simple regression coefficient and multiple regression coefficient will not be equal as the effect of x_2 is ignored by simple regression.

Let us look at the two points when the coefficients will be equal.

- When the coefficient of x_2 is truly zero/does not belong in the model.
- When within the sample x_1 and x_2 are not correlated.

NOTES

It is possible to define the same three sums of squares (SST, SSE and SSR) even in multiple regression. R^2 is the ratio of the explained sum of squares (SSE) to the total sum of squares (SST). The correlation is now not simple though it even now possesses the interpretation of a squared simple correlation coefficient, which is the correlation between y and $\hat{y}$, $r_{\hat{y}y}$.

There will never be a decrease in R^2. It never decreases if there is addition of an explanatory variable to a regression. So, it is possibly to arbitrarily increase the regression R^2 with adding variables, even such variables that are unimportant.

One can fit a model through the origin, suppressing the constant term. But in such a case, several of the mentioned properties will not be there. For example, there is no zero sample average for the least squares residuals (e_is) and there is possibility of a negative R^2 from this type of equation.

In case the population intercept β_0 is not the same as zero, there will be bias in the slope coefficients that is computed via a regression through the origin. So, oftentimes, there is the inclusion of an intercept, and the data is allowed to decide if it should be zero.

OLS Estimators' Expected Value

Let us look at the statistical properties of the OLS estimators associated with the various parameters of the population regression function. It is at (4.3) that the population model has been displayed.

Consider that there is a random sample of size n based upon the model's variables. The multivariate analogue for the assumption regarding the error process will be:

$$E(u \mid x_1, x_2, ..., x_k) = 0 \qquad (4.10)$$

taking the error process as being independent of the distributions. Such an assumption is not valid if there was a misspecification of the model.

There will be a bias also if there is an important separate factor needed to be incorporated with the model. In case that factor gets correlated with the regressors which have been included, there will be a bias in their coefficients.

There is need for making an additional assumption for multiple regression that has several independent variables, with respect to their measured values. Let us look at two propositions in this context.

First Proposition

For the sample, no independent x variable should be expressed as an exact linear relation of the others (including a vector of 1s).

Each multiple regression which has a constant term will be taken as having a variable $x_{0i} = 1 \ \forall i$. According to this proposition, all the other explanatory variables need to possess nonzero sample variance. The proposition further specifies

NOTES

that within the sample there does not exist any perfect collinearity, or multicollinearity. In case one x could be expressed as a linear combination of the other x variables, there would be a violation of this assumption. In case there is perfect collinearity in the regressor matrix, the OLS estimates cannot be computed; mathematically, they do not exist.

Every regressor that is added to a multiple regression has to have information at the margin. It needs to provide some previously unknown information about y.

It is important to understand that this particular proposition is not an assumption about the population model but rather an implication of the sample data being used. Further, this will be valid only in the case of linear relations among the explanatory variables.

Based on the four assumptions (absence of perfect collinearity, zero conditional mean of the u process, random sample and population model), it is possible to show that there is no bias in the population parameters' OLS estimators.

$$Eb_j = \beta_j, \ \ j = 0, ..., k \tag{4.11}$$

Even if the model is misspecified with the inclusion of irrelevant explanatory variables it will not harm the estimates. Unbiased estimates will be yielded by the regression for all coefficients, also for such variables' coefficients that are zero in the population. Further improvement can be attained through removing such variables, since including them in the regression consumes degrees of freedom.

In case the model is under-specified due to the exclusion of some relevant explanatory variable, a problem will occur. Let us look at how serious it can be.

Consider the population model to be as given below:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{4.12}$$

In this we overlook the importance of x_2, and look upon the relationship as being:

$$y = \beta_0 + \beta_1 x_1 + u \tag{4.13}$$

The following displays what the resultant consequences will be:

$$Eb_1 = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)\, x_{i2}}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2} \tag{4.14}$$

This will cause the OLS coefficient b_1 to become biased when the second term is there. When β_2 is nonzero (as is by assumption) and fraction is non zero, the term will also be nonzero. The fraction is no more than a simple regression coefficient in the auxiliary regression of x_2 on x_1. In case of the regressors being correlated with each other, its regression coefficient becomes nonzero, with its magnitude being related to the strength of the correlation (and the units of the variables).

NOTES

Consider the auxiliary regression to be as follows:

$$x_1 = d_0 + d_1 x_2 + u \tag{4.15}$$

where $d_1 > 0$; this makes x_1 and x_2 definitely correlated. The bias can then be depicted as follows:

$$Eb_1 - \beta_1 = \beta_2 d_1 \tag{4.16}$$

with its magnitude and sign dependent on the relation between y and x_2 as well as on the inter-relation amongst the explanatory variables. In case no such relationship exists, there will be an unbiased b_1. Yet, in every other case, bias will exist in the estimation of the underspecified model. In equation (4.16), in case the left side is positive, b_1 will be taken as having an upward bias and there would be an extremely large OLS. In case the left side was negative, there would be a downward bias. In case of the OLS coefficient being nearer zero than to the population coefficient, it would be considered as being attenuated or 'biased toward zero'.

Potential bias is difficult to evaluate with multiple regression, in which the population relationship involves k variables of which only certain are included, such as k - 1.

Generally, each OLS coefficient of an underspecified model will show bias in such a situation unless the variable that has been omitted is uncorrelated with every included regressor. Generally, as a rule asymmetric nature of specification error can be taken away.

Variance of the OLS estimators

Here is the assumption of homoskedasticity, in the context of the k-variable regression model:

$$Var\,(u \mid x_1, x_2, ..., x_k) = \sigma^2 \tag{4.17}$$

In case of the satisfaction of the assumption, error variance will be exactly the same for every combination of the explanatory variables. In case it is not satisfied (is violated), then it is considered that the errors are heteroskedastic. While the OLS estimates will still be unbiased, the estimates of their variances will not be. In the light of the four previous assumptions and this particular one, it is possible to derive the sampling variances/precision, of the OLS slope estimators:

$$Var\left(b_j\right) = \frac{\sigma^2}{SST_j\left(1 - R_j^2\right)},\ j = 1, ..., k \tag{4.18}$$

In (4.18), SST_j is the total variation in x_j about its mean, R_j^2 is the R^2 from an auxiliary regression from regressing x_j on all other x variables, including the

NOTES

constant term. This is a formula applicable to simple regression. For a specific OLS slope estimate to be more precise, there should be an increase in the variation in the associated x variable, making it larger.

When there is perfect collinearity, $R_j^2 = 1$, the sampling variance goes to infinity. In case R_j^2 is extremely tiny, large marginal contributions are made by this variable towards the equation, and it could help to calculate a comparatively more precise estimate of its coefficient. In case R_j^2 is rather large, coefficient's precision stays low as it will not be easy to 'partial out' the effect of variable *j* on *y* from the effects of the other explanatory variables. Yet it must be remembered that assuming that there is no perfect collinearity does not preclude R_j^2 from being close to unity - it just specifies that it is less than unity.

To make (18) operational, the unknown population parameter σ^2 has to be replaced with a consistent estimate:

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{(n-(k+1))} = \frac{\sum_{i=1}^{n} e_i^2}{(n-k-1)} \quad (4.19)$$

In the above, the numerator is just *SSR*, and the denominator is the sample size, less the number of estimated parameters: the constant and *k* slopes.

The additional slope parameters need to be accounted for. This implies that it is not possible to estimate a *k*-variable regression model without a sample of size which is minimum (k+1). For multiple regression, the degrees of freedom shall be positive with k slopes and an intercept, $n > (k + 1)$.

Positive square root of s^2 is called standard error of regression (SER). The units of SER are same as of the dependent variable. It is the numerator of the estimated standard errors of the OLS coefficients. Oftentimes, SER's magnitude is compared with the dependent variable's mean to determine the ability of the regression for explaining the data.

When there is heteroskedasticity, the estimate of s^2 as shown above, remains biased. Similarly, there will be a bias in the estimates of coefficients' standard errors, as they are dependent on s^2.

Example 4.1: The following information is given in the model:

$y_i = \beta_0 + \beta_1 x_i + u_i$

										Total
x_1	1	2	3	4	5	6	7	8	9	10	55
y_1	3.50	1.22	5.86	5.20	10.10	2.38	1.77	11.16	2.71	4.20	48.1

Study the result of applying OLS to the data for estimating the relation.

NOTES

Solution:

x	yi	Σx_i^2	Σy_i^2	$x_i - \overline{x}$ x_i	x_i^2 $(x_i - \overline{x}_2)^2$	y_i $y_i - \overline{y}_i$	y_i^2	$x_i y_i$
1	3.50	1	12.25	–4.5	20.25	–1.31	1.716	+5.895
2	1.22	4	1.49	–3.5	12.25	–3.59	12.898	+12.565
3	5.86	9	34.34	–2.5	6.25	+1.05	1.103	–2.625
4	5.20	16	27.04	–1.5	2.25	+0.39	0.152	–0.585
5	10.10	25	102.01	–0.5	0.25	5.29	27.984	–2.645
6	2.38	36	5.664	–1.5	0.25	–2.43	5.905	–1.215
7	1.77	49	3.133	1.5	2.25	–3.04	9.242	–4.56
8	11.16	64	124.546	2.5	6.25	6.35	40.323	+15.875
9	2.71	81	7.344	3.5	12.25	–2.1	4.41	–7.35
10	4.20	100	17.64	4.5	20.25	–0.61	0.372	–2.745
55	48.1	385	335.4		82.50		104.095	+34.335
								–21.725
								+12.61

$\overline{x} = \dfrac{55}{10}$ $\quad \overline{y} = \dfrac{48.1}{10}$ $\quad \Sigma x_1 = 385$ $\quad \Sigma x_i y_i \;\; 12.62$ $\quad \Sigma y_i^2 \;\; 104.095$

$\overline{x} = 5.5$ $\quad \overline{y} = 4.81$ $\quad \Sigma y_i \;\; 335.455$ $\quad \Sigma x_i^2 \;\; 82.50$

$$\beta_1 = \frac{\Sigma x_i y_i}{\Sigma x_i^2} = \frac{12.62}{82.5} = 0.1529 \text{ or } .153$$

$$\begin{aligned} \beta_0 &= \overline{y} - \beta_1 \overline{x} \\ &= 4.81 - (.153)\,(5.5) \\ &= 4.81 - .8415 \\ &= 3.9685 \text{ or } 3.969 \end{aligned}$$

Therefore $\quad y = \beta_0 + \beta_1 x_1 + y$

Example 4.2: Following table contains observations on the expenditure on clothing (y) total expenditure (x_2) and price of clothing (x_3) we fit a linear regression to there observations and test the overall goodness of fit. (R^2) as well as the statistical reliability of the estimates β_2 and β_3.

NOTES

Table 4.1

n	y Expon Clothing	Total Expenditure x_2	Price of Clothing x_3
1	3.5	15	16.0
2	4.3	20	13.0
3	5.0	30	10.0
4	6.0	42	7.0
5	7.0	50	7.0
6	9.0	54	5.0
7	8.0	65	4.0
8	10.0	72	3.0
9	12.0	85	3.5
10	14.0	90	2.0
Total	78.8	523	70.5

Solution:

$$y_i = \beta + \beta_2 x_{2i} + \beta_3 x_{3i} + u_1$$

$$\beta_1 = \bar{y} - \beta_2 \bar{x}_2 + \beta_3 \bar{x}_3$$

$$\beta_2 = \frac{\Sigma x_3 y \cdot \Sigma x_3^2 - \Sigma x_2 x_3 \cdot \Sigma x_3 y}{\Sigma x_2^2 \, \Sigma x_3^2 - (\Sigma x_2 x_3)^2}$$

$$\beta_3 = \frac{\Sigma x_3 y \cdot \Sigma x_2^2 - \Sigma x_2 x_3 \cdot \Sigma x_2 y}{\Sigma x_2^2 \, \Sigma x_3^2 - (\Sigma x_2 x_2)^2}$$

n	y_1	x_2	x_3	y_i $y_i - \bar{y}$	y_i^2 $(y_i - \bar{y})^2$	x_{2i} $(x_2 - \bar{x})$	x_2^2	x_3 $(x_3 - \bar{x})$	x_3^2	$y_i x_2$	$y_i x_3$	$x_2 x_3$
1	3.5	15	16.0	–4.38	19.18	–37.3	1391.29	8.95	80.10	163.37	–39.20	–333.83
2	4.3	20	13.0	–3.58	12.81	–32.3	1043.29	5.95	35.40	115.63	–21.30	–192.18
3	5.0	30	10.0	–2.88	8.29	–22.3	497.29	2.95	8.70	64.22	–8.49	–65.78
4	6.0	42	7.0	–1.88	3.53	–10.3	106.09	–0.05	0	19.36	0.09	0.51
5	7.0	50	7.0	-0.88	0.77	–2.3	5.29	–0.05	0	2.02	0.04	0.11
6	9.0	54	5.0	+1.12	1.25	1.7	2.89	–2.05	4.20	1.90	–2.29	–3.48
7	8.0	65	4.0	0.12	.01	12.7	161.29	–3.05	9.30	1.52	–0.36	–38.73
8	10.0	72	3.0	2.12	4.49	19.7	388.09	–4.05	16.40	41.76	–8.58	–79.78
9	12.0	85	3.5	4.12	16.97	32.7	1069.29	–3.55	12.60	134.72	–14.62	–116.08
10	14.0	90	2.0	6.12	37.45	37.7	1421.29	–5.05	25.50	230.72	–30.90	–190.38
78.8	523	70.5	0	104.75	0	6086.10	0	192.00	775.22	(–)125.60	(–)1019.62	
					Σy_i^2	Σx_2	Σx_{2i}^2	Σx_{3i}	Σx_{3i}^2	$\Sigma y_i x_{2i}$	$\Sigma y_i x_{3i}$	$\Sigma y_{2i} x_{3i}$

$\overline{y} = 78.8/10 \quad \overline{x}_2 = 523/10 \quad \overline{x}_3 = 70.5/10 \quad \Sigma y_i^2 = 104.75 \quad \Sigma x_{2i}^2 = 6085.10$

$\overline{y} = 7.88 \quad \overline{x}_2 = 52.3 \quad 7.05 \quad \Sigma y_i x_2 = 775.22 \quad \Sigma y_i x_3 = -125.6$

NOTES

$\Sigma x_{3i}^2 = 192.20$

$\Sigma x_2 x_3 = -1019.62$

$$\beta_2 = \frac{\Sigma x_2 y \cdot \Sigma x_3^2 - \Sigma x_2 x_3 \cdot \Sigma x_3 y}{\Sigma x_2^2 \Sigma x_3^2 - (\Sigma x_2 x_2)^2}$$

$$= \frac{(775.22)(192.2) - (-1019.62)(-125.60)}{(6086.1)(192.2) - (-1019.62)^2}$$

$$= \frac{148997.284 - (+128064.272)}{1169748.42 - 1039624.94}$$

$$= \frac{20933.012}{130123.48} = .1608$$

$$\beta_3 = \frac{\Sigma x_3 y_i \cdot \Sigma x_2^2 - \Sigma x_2 x_3 \cdot \Sigma x_2 y}{\Sigma x_2^2 \Sigma x_3^2 - (\Sigma x_2 x_3)^2}$$

$$= \frac{(-125.6)(6086.1) - (-1019.62)(775.22)}{(6036.1)(192.2) - (-1019.62)^2}$$

$$= \frac{-764414.6 - (-790429.816)}{1169748.42 - 1039624.94}$$

$$= \frac{26015.6564}{130123.48} = 0.1999$$

Therefore

$$\beta_1 = y - \beta_2 \overline{x}_2 - \beta_3 \overline{x}_3$$

$$\beta_1 = 7.88 - (.1608)(52.3) - (.200)(7.05)$$

$$= 7.88 - 8.4098 - 1.41$$

$$= -1.9398 \text{ or } -1.94$$

$$R^2 = \frac{\beta_2 \Sigma x_i y_i + \beta_3 \Sigma x_3 y_i}{\Sigma y_i}$$

$$\frac{(.1608)(775.22) + (.200)(-125.60)}{104.75}$$

$$= \frac{124.655 + (-25.12)}{104.75}$$

$$= \frac{99.535}{104.75} = .95$$

$$R^2 = .95 \text{ or}$$

$$R^2 = \frac{Ess}{Tss} = 1 - \frac{\Sigma e_i}{\Sigma y_i^2}$$

(iii) For Estimation of the Standard error of β_2 .. β_2 we need to estimate of σ_u^2

$$\sigma_u^2 = \frac{\Sigma e_i^2}{n-2}$$

$$\Sigma e_i^2 = \Sigma y_i^2 (1 - R^2)$$

$$\Sigma e_i^2 = 104.75\ (1 - .95)$$

$$= 104.75(.05)$$

$$= 5.23$$

Therefore

$$\sigma_u^2 = \frac{\Sigma e_i^2}{n-2} = \frac{5.23}{10-3} = \frac{5.23}{7} = .74$$

$$\text{Var}\ (\beta_2) = \frac{\delta_u^2 \Sigma x_3^2}{\Sigma x_2^2 \Sigma x_3^2 - (\Sigma x_2 x_3)^2} = \frac{(.74)(192.2)}{(6068.1)(192.2) - (-1019.62)^2}$$

$$= \frac{142.228}{130123.48} = .0011$$

$$\text{V}\ \ \text{SE}\ (\beta_2)\ \sqrt{\text{Var}\,\beta_2} = \sqrt{.0011} = .03306 \text{ or } 0.331$$

$$\text{Var}\ \beta_3 = \frac{\delta_u^2 \Sigma x_2^2}{\Sigma x_2^2 \Sigma x_3^2 - (\Sigma x_2 x_3)^2} = \frac{(.74)6086.1}{(6068.1)(192.2) - (-1019.62)^2}$$

$$\frac{4503.714}{130123.48} = .0346$$

$$\text{SE}(\beta_3) = \sqrt{\text{Var}\,\beta_3} = \sqrt{.0346} = .186$$

4.3 ESTIMATION OF PARAMETERS

Economic, Social and Scientific variables are often times connected by linear associations with the assumption that the model is linear in parameters. In the field of econometrics and statistical modeling, regression analysis is a conceptual process for ascertaining the functional relationship that exists among variables.

Several models used in stock assessment were analysed, the respective parameters having been defined. In the corresponding exercises, it was not necessary to estimate the values of the parameters because they were given. In this unit, several methods of estimating parameters will be analysed. In order to estimate the parameters, it is necessary to know the sampling theory and statistical inference.

This manual will use one of the general methods most commonly used in the estimation of parameters - the least squares method. In many cases this method uses iterative processes, which require the adoption of initial values. Therefore, particular methods will also be presented, which obtain estimates close to the real values of the parameters. In many situations, these initial estimates also have a practical interest. These methods will be illustrated with the estimation of the growth parameters and the S-R stock-recruitment relation.

The least squares method is presented under the forms of Simple linear Regression, multiple linear model and non-linear models (method of Gauss-Newton).

4.3.1 Simple Linear Regression - Least Squares Method Model

Consider the following variables and parameters:

Response or Dependent Variable= Y

Auxiliary or Independent Variable= X

Parameters= A, B

The response variable is linear with the parameters

Y = A+BX

Aim

The aim of LSM method under SLR is to estimate the parameters of the model, based on the observed pairs of values and applying a certain criterium function (the observed pairs of values are constituted by selected values of the auxiliary variable and by the corresponding observed values of the response variable), that is:

Observed values- x_i and y_i for each pair i, where i=1,2,...,i,...*n*

Values to be estimated A and B and $(Y_1, Y_2, ..., Y_i, ..., Y_n)$ for the *n* observed pairs of values

Estimates values : $\hat{A}$ and $\hat{B}$ (or a and b) and $(\hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_i ..., \hat{Y}_n)$

Object function (or criterium function)

$$\Phi = \sum_{i=1}^{n} (y_i - Y_i)^2$$

Estimation Method

In the least squares method the estimators are the values of A and B which minimize the object function. Thus, one has to calculate the derivatives ∂Φ/∂A e ∂Φ/∂B, equate them to zero and solve the system of equations in A and B.

NOTES

The solution of the system can be presented as:

$\bar{x} = (1/n).\sum x$ $\bar{y} = (1/n).\sum y$

$Sxx = \sum(x - \bar{x})(x - \bar{x})$ $Sxy = \sum(x - \bar{x})(y - \bar{y})$

b = Sxy/Sxx $a = \bar{y} - b.\bar{x}$

Notice that the observed values y, for the same set of selected values of X, depend on the collected sample. For this reason, the problem of the simple linear regression is usually presented in the form:

$y = A + BX + \varepsilon$

Where ε is a random variable with *expected value* equal to zero and *variance* equal to σ^2.

So, the expected value of y will be Y or A+BX and the variance of y will be equal to the variance of ε.

The terms deviation and residual will be used in the following ways:

Deviation is the difference between $y_{observed}$ and y_{mean} ($\bar{y}$), i.e., deviation = ($y-\bar{y}$)

While

Residual is the difference between $y_{observed}$ and $Y_{estimated}$ ($\hat{Y}_i$), i.e., residual = $y_i - \hat{Y}_i$.

To analyse the adjustment of the model to the observed data, it is necessary to consider the following characteristics:

Sum of Squares of the Residuals

$$SQ_{residual} = \sum(y - \hat{Y})^2$$

This quantity indicates the residual variation of the observed values in relation to the estimated values of the response variable of the model, which can be considered as the variation of the observed values that is not explained by the model.

Sum of Squares of the Deviations of the Estimated Values of the Response Variable of the Model

$$SQ_{model} = \sum(\hat{Y} - \bar{y})^2$$

This quantity indicates the variation of the estimated values of the response variable of the model in relation to its mean that is the *variation* of the response variable explained by the model.

Total Sum of Squares of the Deviations of the Observed Values Equal To

$$SQ_{residual} = \sum \left(y - \bar{y}\right)^2$$

NOTES

This quantity indicates the total variation of the observed values in relation to the mean

It is easy to verify the following relation:

$SQ_{Total} = SQ_{Model} + SQ_{Residual}$

Or

$$1 = \frac{SQ_{model}}{SQ_{total}} + \frac{SQ_{residual}}{SQ_{total}}$$

Or

$1 = r^2 + (1 - r^2)$

Whereas: r^2 (coefficient of determination) is the percentage of the total variation that is explained by the model and

$1-r^2$ is the percentage of the total variation that is not explained by the model.

4.3.2 Multiple Linear Regression - Least Squares Method

Model

Consider the following variables and parameters:

Response or dependent variable = Y

Auxiliary or independent variables = $X_1, X_2, ..., X_j, ..., X_k$

Parameters = $B_1, B_2, ..., B_j, ..., B_k$

The response variable is linear with the parameters

$Y = B_1X_1 + B_2X_2 + ... + B_kX_k = \Sigma B_jX_j$

Aim

The aim of the Leat Squares method under the MLR is to estimate the parameters of the model, based on the observed **n** sets of values and by applying a certain criterium function (the observed sets of values are constituted by selected values of the auxiliary variable and by the corresponding observed values of the response variable), that is:

Observed values $x_{1,i}\, x_{2,i,...}, x_{j,i,..}, x_{k,i}$ and y_i for each set i, where i=1,2,...,i,...n

Values to be estimated $B_1, B_2, ..., B_j, ..., B_k$ et $(Y_1, Y_2, ..., Y_i, ..., Y_n)$

The estimated values can be represented by:

$\hat{B}_1, \hat{B}_2, ..., \hat{B}_j, ..., \hat{B}_k$ (ou $b_1, b_2, ..., b_j, ..., b_k$) et $\hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_i, ..., \hat{Y}_n$

Object Function (or Criterium Function)

$$\Phi = \sum_{i=1}^{n} (y_i - Y_i)^2$$

Estimation Method

In the *least squares method* the estimators are the values of B_j which minimize the object function.

As with the simple linear model, the procedure of minimization requires equating the partial derivatives of Φ to zero in order to each parameter, B_j, where j=1, 2,..., k. The system is preferably solved using matrix calculus.

Matrix Version

Matrix $X_{(n,k)}$ = Matrix of the n observed values of each of the k auxiliary variables

Vector $y_{(n,1)}$ = Vector of the n observed values of the response variable

Vector $Y_{(n,1)}$ = Vector of the values of the response variable given by the model (unknown)

Vector $B_{(k,1)}$ = Vector of the parameters

Vector $\hat{B}$ or $b_{(k,1)}$ = Vector of the estimators of the parameters

Model

$Y_{(n,1)} = X_{(n,k)}. B_{(k,1)}$ ou $Y=X.B+\varepsilon$

Object Function

$\Phi_{(1,1)} = (y-Y)^T.(y-Y)$ ou $\Phi_{(1,1)} = (y-X.B)^T.(y-X.B)$

To calculate the least squares estimators it will suffice to put the derivative dΦ/dB of Φ in order to vector B, equal to zero. dΦ/dB is a vector with components $\partial\Phi/\partial B_1, \partial\Phi/\partial B_2,..., \partial\Phi/\partial B_k$. Thus:

$d\Phi/dB_{(k,1)} = -2.X^T.(y-X.B) = 0$

Or $X^Ty - (X^T.X). B = 0$

and $b = \hat{B} = (X^T.X)^{-1.} X^Ty$

The results can be written as:

$b_{(k,1)} = (X^T.X)^{-1}.X^Ty$

$\hat{Y}_{(n,1)} = X.b$ or $\hat{Y}_{(n-1)} = X (X^T.X)^{-1}.X^T y$

$\text{Residuals}_{(n,1)} = (y- \hat{Y})$

NOTES

Comments

In statistical analysis it is convenient to write the estimators and the sums of the squares using idempotent matrices. Then the idempotent matrices L, (I - L) and (I - M) with $L_{(n,n)} = X (X^{T.} X)^{-1.} X^T$, I = unity matrix and $M_{(n,n)}$ = $mean_{(n,1)}$ matrix = 1/n [1] where [1] is a matrix with all its elements equal to one, are used.

It is also important to consider the sampling distributions of the estimators assuming that the variables ε_i are independent and have a normal distribution.

A summary of the main properties of the expected value and variance of the estimators is presented:

	$E[c_1+c_2.u] = c_1+c_2.E[u]$	$V[c_1+c_2.u] = c_2.V[u].c_2^T$
1 -	Random variable, ε	ε_n (independent)
	Expected value of ε	E[ε] = 0.
	Variance of ε	$V[\varepsilon]_{(n,n)} = E[\varepsilon.\varepsilon^T]=I.\ \sigma^2$
2 -	Observed response variable y	y = Y+ε
	Expected value of y	E[y] = Y = X.B.
	Variance of y	$V[y]_{(n,n)} = V[\varepsilon]_{(n,n)} = I.\ \sigma^2$
3 -	Estimator of B	$\hat{B} = (X^T.X)^{-1}.X^T.y$
	Expected value of $\hat{B}$	$E[\hat{B}] = B$
	Variance of $\hat{B}$	$V[\hat{B}]_{(k,k)} = (X^T.X)^{-1}.\ \sigma^2$
4 -	Estimator of Y of the model	$\hat{Y} = X.\ \hat{B} = L.y$
	Expected value of $\hat{Y}$	$E[\hat{Y}] = Y.$
	Variance of $\hat{Y}$	$V[\hat{Y}] = L.\ \sigma^2$
5 -	Residual e	$e = y - \hat{Y} = (I-L).y$
	Expected value of e	E[e] = 0
	Variance of e	$V[e] = (I-L).\ \sigma^2$

Sum of squares

Residual Sum of squares = SQ $residual_{(1.1)} = (y-\hat{Y})^T(y-\hat{Y}) = y^T (I-L)y$

This quantity indicates the residual variation of the observed values in relation to the estimated values of the model, i.e., the variation not explained by the model.

Sum of squares of the deviation of the model = SQ $model_{(1.1)} = (\hat{Y}-\bar{y})^T (\hat{Y}-\bar{y}) = y^T (L-M)y$

This quantity indicates the variation of the estimated response values of the model in relation to the mean, that is, the variation explained by the model.

Total Sum of the squares of the deviations = SQ total$_{(1.1)}$ = $(y-\bar{y})^T(y-\bar{y})$ = y^T (I-M) y

This quantity indicates the total variation of the observed values in relation to the mean.

It is easy to verify the following relation:

$SQ_{total} = SQ_{model} + SQ_{residual}$ or

$$1 = \frac{SQ_{model}}{SQ_{total}} + \frac{SQ_{residual}}{SQ_{total}}$$

Or $1 = R^2 + (1 - R^2)$

Where,

R^2 is the percentage of the total variation that is explained by the model. In matrix terms it will be:

$R^2 = [y^T (L - M) y]. [(y^T(I - M)y]^{-1}$

$1-R^2$ is the percentage of the total variation that is not explained by the model.

The ranks of the matrices (I-L), (I-M) and (L-M) respectively equal to (n-k), (n-1) and (k-1), are the degrees of freedom associated with the respective sums of squares.

4.3.3 Non-Linear Model - Method of Gauss-Newton - Least Squares Method

Model

Consider the following variables and parameters:

Response or dependent variable = Y

Auxiliary or independent variable = X

Parameters = $B_1, B_2, ..., B_j, ..., B_k$

The response variable is non-linear with the parameters

Y = f(X; B) where B is a vector with the components B_1, B2,...,B_j,...,B_k

Aim

The aim of the method is to estimate the parameters of the model, based on the n observed pairs of values and by applying a certain criterium function (the observed sets of values are constituted by selected values of the auxiliary variable and by the corresponding observed values of the response variable), that is:

Observed values x_i and y_i for each pair i, where i=1, 2,...,i,...n

Values to be estimated $B_1, B_2, ..., B_j, .., B_k$ and $(Y_1, Y_2, ..., Y_i, ..., Y_n)$ form the n pairs of observed values.

(Estimates = $\hat{B}_1, \hat{B}_2, ..., \hat{B}_j, ..., \hat{B}_k$ or b_1, b2,...,b_j,...,b_k and $\hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_i, ..., \hat{Y}_n$)

Object Function or Criterium Function

$$\Phi = \sum_{i=1}^{n} (y_i . Y_i)^2$$

NOTES

Estimation Criterium

The estimators will be the values of B_j for which the object function is minimum.

(This criterium is called the least squares method).

Matrix Version

It is convenient to present the problem using matrices.

So

Vector $X_{(n,1)}$ = Vector of the observed values of the auxiliary variable

Vector $y_{(n,1)}$ = Vector of the observed values of the response variable

Vector $Y_{(n,1)}$ = Vector of the values of the response variable given by the model

Vector $B_{(k,1)}$ = Vector of the parameters

Vector $b_{(k,1)}$ = Vector of the estimators of the parameters

Model

$Y_{(n,\ 1)} = f(X; B)$

Object Function

$\Phi_{(1,1)} = (y-Y)^T.(y-Y)$

In the case of the non-linear model, it is not easy to solve the system of equations resulting from equating the derivative of the function Φ in order to the vector B, to zero. Estimation by the least squares method can, based on the Taylor series expansion of function Y, use iterative methods.

Revision of the Taylor Series Expansion of a Function

Here is an example of the expansion of a function in the Taylor series in the case of a function with one variable.

The approximation of Taylor means to expand a function Y = f(x) around a selected point, x_0, in a power series of x:

$Y = f(x) = f(x_0) + (x-x_0).f'(x_0)/1! + (x-x_0)^2 f''(x_0)/2! + ... + (x- x_0)^i f^{(i)}(x_0)/i! + ...$

Where

$f^{(i)}(x_0) = i^{th}$ derivatives of f(x) in order to x, at the point x_0.

The expansion can be approximated to the desired power of x. When the expansion is approximated to the power 1 it is called a linear approximation, that is,

$Y \cong f(x_0) + (x-x_0).f'(x_0)$

NOTES

The Taylor expansion can be applied to functions with more than one variable. For example, for a function $Y = f(x_1, x_2)$ of two variables, the linear expansion would be:

$$Y \approx f(x_{1(0)}, x_{2(0)}) + (x_1 - x_{1(0)}).\frac{\delta f(x_{1(0)}, x_{2(0)})}{\delta x_1} + (x_2 - x_{2(0)}).\frac{\delta f(x_{1(0)}, x_{2(0)})}{\delta x_2}$$

Which may be written, in matrix notation, as

$Y = Y_{(0)} + A_{(0)}.(x-x_{(0)})$

where $Y_{(0)}$ is the value of the function at the point $x_{(0)}$,with components $x_{1(0)}$ and $x_{2(0)}$,and $A_{(0)}$ is the matrix of derivatives whose elements are equal to the partial derivatives of $f(x_1,x_2)$ in order to x_1,x_2 at the point $(x_{1(0)}, x_{2(0)})$.

To estimate the parameters, the Taylor series expansion of function Y is made in order to the parameters B and not to the vector X.

For example, the linear expansion of Y = f(x,B) in $B_1, B_2,..., B_k$, would be:

$Y = f(x;B) = f(x; B_{(0)}) + (B_1-B_{1(0)})\ f/B_1\ (x;B_{(0)}) +..... +$

$(B_2-B_{2(0)})f/B_2\ (x;B_{(0)}) +...... +.........+ (B_k-B_{k(0)})\ f/B_k\ (x;B_{(0)})$

Or, in matrix notation, it would be:

$Y_{(n,1)} = Y_{(0)\ (n,1)} + A_{(0)\ (n,k).}\ \Delta B_{(0)\ (k,1)}$

Where

A = matrix of order (n,k) of the partial derivatives of the matrix f(x;B) in order to the vector B at the point $B_{(0)}$ and

$\Delta B_{(0)}$ = vector $(B - B_{(0)})$.

Then, the object function will be:

$\Phi = (y-Y)^T.(y-Y) = (y-Y_{(0)} - A_{(0)}.\ \Delta B_{(0)})^T(y-Y_{(0)} - A_{(0).}\ \Delta B_{(0)})$

To obtain the minimum of this function it is more convenient to differentiate Φ in order to the vector ΔB *than in relation to vector* B *and put it equal to zero*. Thus:

$$0 = -2(A_{(0)})^T(y - Y_{(0)} - A_{(0)}.\Delta B_{(0)}) = -2A_{(0)}^T(y - Y_{(0)}) + 2A_{(0)}^T A_{(0)}.\Delta B_{(0)}$$

Or

$$A_{(0)}^T A_{(0)}.\Delta B_{(0)} = A_{(0)}^T(y - Y_{(0)})$$

Therefore:

$$\Delta B_{(0)} = \left(A_{(0)}^T.A_{(0)}\right)^{-1}.A_{(0)}^T.(y - Y_{(0)})$$

If $\Delta B_{(0)}$ is 'Equal to Zero' then the estimate of B is equal to $B_{(0)}$.

(In practice, when we say equal to zero in this process, we really mean smaller than the approximation vector one has to define beforehand).

NOTES

If $\Delta B_{(0)}$ is not "equal to zero" then the vector $B_{(0)}$ will be replaced by:

$B_{(1)} = B_{(0)} + \Delta B_{(0)}$

And the process will be repeated, that is, there will be another iteration with $B_{(0)}$ replaced by $B_{(1)}$ (and $A_{(0)}$ replaced by $A_{(1)}$). The iterative process will go on until the convergence at the desired level of approximation is reached.

Comments

1. It is not guaranteed that the process always converges. Sometimes it does not, some other times it is too slow (even for computers!) and some other times it converges to another limit!!
2. The above described method is the Gauss-Newton method which is the basis of many other methods. Some of those methods introduce modifications in order to obtain a faster convergence like the Marquardt method (1963), which is frequently used in fisheries research. Other methods use the second order Taylor expansion (Newton-Raphson method), looking for a better approximation. Some others, combine the two modifications.
3. These methods need the calculation of the derivatives of the functions. Some computer programs require the introduction of the mathematical expressions of the derivatives, while others use sub-routines with numerical approximations of the derivatives.
4. In fisheries research, there are methods to calculate the initial values of the parameters, for example in growth, mortality, and selectivity or maturity analyses.
5. It is important to point out that the convergence of the iterative methods is faster and more likely to approach the true limit when the initial value of the vector $B_{(0)}$ is close to the real value.

4.3.4 Estimation of Growth Parameters

The least squares method (non-linear regression) allows the estimation of the parameters K, L_∞ and t_o of the individual growth equations.

The starting values of K, L_∞ and t_0 for the iterative process of estimation can be obtained by simple linear regression using the following methods:

Ford-Walford (1933-1946) and Gulland and Holt (1959) Methods

The Ford-Walford and Gulland and Holt expressions, are already in their linear form, allowing the estimation of K and $L_{,,}$ with methods of simple linear regression on observed L_i and T_i. The Gulland and Holt expression allows the estimation of K and $L_{,,}$ even when the intervals of time T_i are not constant. In this case, it is convenient to re-write the expression as:

$$\Delta L / T_i = K.L_\infty - K.\overline{L}$$

NOTES

Stamatopoulos and Caddy Method (1989)

These authors also present a method to estimate K, L_∞ and t_o (or L_o) using the simple linear regression. In this case the von Bertalanffy equation should be expressed as a linear relation of L_t against e^{-Kt}.

Consider n pairs of values t_i, L_i where t_i is the age and L_i the length of the individual i where i=1, 2,...., n.

The von Bertalanffy equation, in its general form is (as previously seen):

$L_\infty - L_t = (L_\infty - L_a). e^{-K(t-ta)}$

It can be written as:

$L_t = L_\infty - (L_\infty - L_a). e^{+Kta}. e^{-Kt}$

The equation has the simple linear form, y = a + bx, where:

$y = L_t \quad a = L_\infty \quad b = -(L_\infty - L_a). e^{+Kta}$

$x = e^{-Kt}$

If one takes $L_a = 0$, then $t_a = t_o$, but, if one considers $t_a = 0$, then $L_a = L_o$.

The parameters to estimate from a and b will be L_∞, t_o or L_o.

The authors propose adopting an initial value $K_{(0)}$, of K, and estimating $a_{(0)}$, $b_{(0)}$ and $r^2_{(0)}$ by simple linear regression between y (= L_t) and x(=$e^{k}_{(0)}$). The procedure may be repeated for several values of K, that is, $K_{(1)}$ $K_{(2)}$,....K_n One can then adopt the regression that results in the larger value of r^2, to which K_{max}, a_{max} and b_{max} correspond. From the values of a_{max}, b_{max} and K_{max} one can obtain the values of the remaining parameters.

One practical process towards finding K_{max} can be following Steps:

(i) To select two extreme values of K which include the required value, for example K= 0 and K=2 (for practical difficulties, use K = 0.00001 instead of K = 0).

(ii) Calculate the 10 regressions for equally-spaced values of K between those two values in regular intervals.

(iii) The corresponding 10 values of r^2 will allow one to select two new values of K which determine another interval, smaller than the one in **(i),** containing another maximum value of r^2.

(iv) The Steps **(ii)** and **(iii)** can be repeated until an interval of values of K with the desired approximation is obtained.

4.4 PROPERTIES OF OLS ESTIMATORS

Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable. Multiple Liner Regression (MLR) is used extensively in econometrics and financial inference.

Principle of Ordinary Least Squares (OLS)

Let B be the set of all possible vectors β. If there is no further information, the B is k -dimensional real Euclidean space. The object is to find a vector b' = (b_1, b_2,…..,b_k) = from B that minimizes the sum of squared deviations of, i ε'_is, i.e.,

For given y and X. A minimum will always exist as S (β) is a real-valued, convex and differentiable function. Write

$S(\beta) = y'y + \beta'X'X\beta - 2\beta'X'y$

Differentiate S (β) with respect to β

$$\frac{\partial S(\beta)}{\partial \beta} = 2X'X\beta - 2X'y$$

$$\frac{\partial^2 S(\beta)}{\partial \beta^2} = 2X'X \quad \text{(At least non-negative definite)}$$

Where the following result is used

Result: If f (*z*) = Z′AZ = is a quadratic form, Z is a m×1 vector and A is any m× m symmetric matrix

Then

$$\frac{\partial}{\partial z} F(z) = 2Az$$

Since it is assumed that rank (X) = k (full rank), then X′ X is a positive definite and unique solution of the normal equation is

$b = (X'X)^{-1} X'y$

Which is termed as Ordinary Least Squares Estimator (OLSE) of β .

Since $\frac{\partial^2 S(\beta)}{\partial \beta^2}$ is at least non-negative definite, so b minimize S (β).

In case, X is not of full rank, then

$B = (X'X)^{-}X'y + [I - (X'X)^{-}X'X]\,\omega$

Where (X′ X) is the generalised inverse of X′X and ω is an arbitrary vector. The generalized inverse $(X'X)^{-}$ of X′ X satisfies

$X'X(X'X)^{-}X'X = X'X$

$X(X'X)^{-}X'X = X$

$X'X(X'X)^{-}X'X = X'$

Theorem

(i) Let $\hat{Y} = Xb$ be the empirical predictor of y. Then $\hat{Y}$ has the same value for all solutions b of $X'Xb = X'y$.

(ii) $S(\beta)$ attains the minimum for any solution of $X'Xb = X'y$

Proof

(i) Let b be any member in

$b = (X'X)^{-}X'y + [I - (X'X)^{-} X'X]\,\omega$

Since $X(X'X)^{-} X'X = X$ so then

$Xb = X(X'X)^{-}X'y + X\,[I - (X'X)^{-} X'X]\,\omega$

$= X(X'X)^{-}X'y$

Which is independent of ω. This implies that $\hat{Y}$ has the same value for all solution b of $X'Xb = X'y$

(ii) Note that for any β ,

$$\begin{aligned}
S(\beta) &= [y - Xb + X(b - \beta)]'\,[y - Xb + X(b - \beta)] \\
&= (y - Xb)'(y - Xb) + (b - \beta)'X'X(b - \beta) + 2(b - \beta)'X'(y - Xb) \\
&= (y - Xb)'(y - Xb) + (b - \beta)'X'X(b - \beta) \\
&\qquad\qquad (\text{Using } X'Xb = X'y) \\
&\geq (y - Xb)'(y - Xb) = S(b) \\
&= y'y - 2y'Xb + b'X'Xb \\
&= y'y - b'X'Xb \\
&= y'y - \hat{Y}\hat{Y}
\end{aligned}$$

Fitted Values

If $\hat{\beta}$ is any estimator of β for the model $y = X\beta + \varepsilon$, then the fitted values are defined as $\hat{Y} = \mathbf{X}\hat{\beta}$ is any estimator of β.

In the case of $\hat{\beta} = b$,

$$\begin{aligned}
\hat{\mathbf{Y}} &= \mathbf{Xb} \\
&= \mathbf{X(X'X)^{-1}X'y} \\
&= \mathbf{Hy}
\end{aligned}$$

Where $H = X(X'X)^{-1}X'$ is termed as Hat Matrix, which is

(i) Symmetric

(ii) Idempotent, (i.e., $HH = H$) and

(iii) $tr\, H\; tr\, \mathbf{X(X'X)^{-1}} = tr\, \mathbf{X'X(X'X)^{-1}}\; tr\, I_k = k$

Residuals

The difference between the observed and fitted values of the study variable is called as residual. It is denoted as

NOTES

$$e = y \sim \hat{Y}$$

$$= y - \hat{Y}$$

$$\mathbf{= y - Xb}$$

$$\mathbf{= y - Hy}$$

$$\mathbf{= (I-H)y}$$

$$= \bar{H}\mathbf{y}$$

Where $\bar{H}$ **= I- H**

Note: $\bar{H}$ **is a symmetric and idempotent matrix**

Properties of OLSE

(i) Estimation Error: The estimation error of b is:

$$b - \beta = \mathbf{(X'X)^{-1}X'}\, y - \beta$$

$$= \mathbf{(X'X)^{-1}X'}\,(X\beta = \varepsilon) - \beta$$

$$= \mathbf{(X'X)^{-1}X'}\varepsilon$$

(ii) Bias

Since X is assumed to be non- stochastic and E (ε) = 0

$$E\,(b-\beta) = \mathbf{(X'X)^{-1}X'}\, E\,(\varepsilon)$$

$$= 0$$

Thus OLSE is an unbiased estimator of β

(iii) Covariance Matrix

The covariance matrix of b is

$$V(b) = E(b-\beta)(b-\beta)'$$
$$= E\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right]$$
$$= (X'X)^{-1}X'E(\varepsilon\varepsilon')X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}X'IX(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}.$$

(iv) Variance

The variance of b can be obtained as the sum of variances of all b_1, b_2,..., b_k which is the trace of covariance matrix of b. Thus

$$
\begin{aligned}
Var(b) &= tr[V(b)] \\
&= \sum_{i=1}^{k} E(b_i - \beta_i)^2 \\
&= \sum_{i=1}^{k} Var(b_i).
\end{aligned}
$$

Estimation of σ^2

The least-squares criterion cannot be used to estimate σ^2 because σ^2 does not appear in S (β). Since E (ε^2_i) = σ^2, so we attempt with residuals e_i to estimate as σ^2 follows:

$$
\begin{aligned}
e &= y - \hat{y} \\
&= y - X(X'X)^{-1}X'y \\
&= [I - X(X'X)^{-1}X']y \\
&= \bar{H}y.
\end{aligned}
$$

Consider the residual sum of squares

$$
\begin{aligned}
SS_{res} &= \sum_{i=1}^{n} e_i^2 \\
&= e'e \\
&= (y - Xb)'(y - Xb) \\
&= y'(I - H)(I - H)y \\
&= y'(I - H)y \\
&= y'\bar{H}y.
\end{aligned}
$$

Also

$$
\begin{aligned}
SS_{res} &= (y - Xb)'(y - Xb) \\
&= y'y - 2b'X'y + b'X'Xb \\
&= y'y - b'X'y \quad (\text{Using } X'Xb = X'y)
\end{aligned}
$$

$$
\begin{aligned}
SS_{res} &= y'\bar{H}y \\
&= (X\beta + \varepsilon)'\bar{H}(X\beta + \varepsilon) \\
&= \varepsilon'\bar{H}\varepsilon \quad (\text{Using } \bar{H}X = 0)
\end{aligned}
$$

Since $\varepsilon \sim N(0, \sigma^2 I)$, so $y \sim N(X\beta, \sigma^2 I)$. Hence $y'\bar{H}y \sim \chi^2(n-k)$

Thus $E[y'\bar{H}y] = (n-k)\sigma^2$

or $E\left[\dfrac{y'\bar{H}y}{n-k}\right] = \sigma^2$

or $E[MS_{res}] = \sigma^2$

NOTES

Where $MS_{res} = \frac{SS_{res}}{n-k}$ is the mean sum of squares due to residual.

Thus an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = MS_{res} = s^2 \text{ (say)}$$

Which is a model-dependent estimator.

Variance of $\hat{Y}$

The variance of $\hat{Y}$ **is**

$$\begin{aligned} V(\hat{y}) &= V(Xb) \\ &= XV(b)X' \\ &= \sigma^2 X(X'X)^{-1}X' \\ &= \sigma^2 H. \end{aligned}$$

Check Your Progress

1. What is multiple regression model?
2. What do you understand by least squares normal equation?
3. Elaborate on the standard error of regression.
4. Explain the aim of least square method model under the SLR.
5. Interpret the estimation method of least squares method.
6. Explain the aim of least squares method under the MLR.
7. Write a short note on Ford – Walford, Gulland and Holt method.
8. Give the uses of MLR.
9. What is Variance of $\hat{Y}$?

4.5 GOODNESS OF FIT

Having developed a regression equation, it is very appropriate to understand how well this regression line fits the observed data.

(i) **Coefficient of determination:** The coefficient of determination, R^2, is a widely used measure of the goodness of fit of a regression line. It measures the extent or strength of the association that exists between the two variables, X and Y. Since this value is based on the sample data points, it is known as the 'Sample Coefficient of determination'. It's value ranges from 0 (poor fit) to 1 (good fit) or (perfect fit). R^2 is based on two kinds of variation:

(i) variation of Y values around the fitted regression line and

(ii) variation of Y around their own mean

Thus, the variation of the Y values around the regression line is given by $\Sigma(Y-\hat{Y})^2$ and around their own mean is given by $\Sigma(Y-\overline{Y})^2$. $\hat{Y} = a + b\hat{X}$ is the regression line.

Thus, $R^2 = \dfrac{\text{'explained variation'}}{\text{total variation.}}$

total variation = explained + unexplained.

R^2 is defined as the proportion of the total variation, i.e., 'explained' and 'unexplained'.

i.e; $\Sigma(Y-\overline{Y})^2 = \Sigma(\hat{Y}-\overline{Y})^2 + \Sigma(Y-\hat{Y})^2$

Thus, $R^2 = \dfrac{\text{total variation} - \text{unexplained variation}}{\text{total variation}}$

$$= 1 - \frac{\Sigma(Y-\hat{Y})^2}{\Sigma(Y-\overline{Y})^2}$$

R^2 gives the proportion of variation in Y that can be accounted for by the variation in X. Also R^2 can be used to measure how well X variable explains Y.

$R^2 \to 1$ shows close correlation between X and Y. A simplified formula for R^2 is

$$R^2 = \frac{\hat{a}\Sigma Y + \hat{b}\Sigma XY - n\overline{Y}^2}{\Sigma Y^2 - n\Sigma\overline{Y}^2}$$

For example, 1. If $r^2 = 0.64 \Rightarrow$ that only 64% of the variation in the relative series has been explained by the subject series and the remaining variation is due to other fallers r^2 is non-negative and it does not tell us the direction of relationship between the two series.

2. When the sample size is small, the estimate of R^2 is positively biased, i.e., R^2 tends to be the higher side. An unbiased estimate for R^2 known as the adjusted coefficient of determination, $\overline{R}^2$ is given by

$$\overline{R}^2 = 1 - \frac{\text{residual variance}}{\text{total variance}}$$

$$= 1 - \frac{\Sigma(Y_i-\hat{Y}_i)^2.(n-1)}{\Sigma(Y_i-\overline{Y})^2.(n-2)}$$

The adjustment factor being $(n-1)/(n-2) > 1$, R^2 will always be less than or equal to $\overline{R}^2$ as the sample size increases, the ratio $\dfrac{(n-1)}{(n-2)} \to 1$ and thus, the difference between R^2 and $\overline{R}^2$ will be reduced.

4.6 R² AND ADJUSTED R²

NOTES

For a simple regression mode, it is possible to explain the variation of one variable with the help of another due to the fact that they have a correlation. In case they were not correlated, no explanatory power would be there in the *X* variable. During regression analysis, there is a clear relation between the correlation coefficient and the coefficient of determination, though there is a slight difference in their interpretation. Moreover, use can be made of the correlation coefficient only between pairs of variables, and the coefficient of determination has the ability to connect a group of variable with the dependent variable.

No information is provided by the correlation coefficient regarding two variables' causal relationship. Let us look at the correlation coefficient in a context of the regression model and know in what circumstances it will be right to interpret the correlation coefficient as a measure of strength of a causal relationship.

In **coefficient of determination**, the average deviation from mean is split into two parts: explained and an unexplained part. So, it is natural to start the derivation of the measure from the deviation from the mean expression, following which, the predicted value obtained from the regression model are introduced. So, for a single individual:

$$Y_i - \bar{Y} = Y_i - \bar{Y} + \hat{Y}_i - \hat{Y}_i = \underbrace{\left(\hat{Y}_i - \bar{Y}\right)}_{\text{Explained}} + \underbrace{\left(Y_i - \hat{Y}_i\right)}_{\text{Unexplained}} \tag{4.20}$$

The deviation from the mean value of Y is explained by employing the regression model. So, we will denote the difference between the expected value (y) and the mean value *(y)* as the mean difference's explained part. The part that remains is the unexplained part. In this manner, decomposing is done for a single observation with respect to the simple mean difference. Now, (4.20) needs to be transformed to such an expression which will hold for the complete sample (for every observation). This is done with squaring and summing over each of the *n* observations:

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(\left(\hat{Y}_i - \bar{Y}\right) - \left(Y_i - \hat{Y}_i\right)\right)^2 = \sum_{i=1}^{n}\left[\left(\hat{Y}_i - \bar{Y}\right)^2 + \left(Y_i - \hat{Y}_i\right)^2 - 2\left(\hat{Y}_i - \bar{Y}\right)\left(Y_i - \hat{Y}_i\right)\right]$$

It can be shown that for the last expression on the right hand side the sum is equal to zero. In that light:

$$\underbrace{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}_{TSS} = \underbrace{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}_{ESS} + \underbrace{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}_{RSS}$$

After the manipulations have been made, there are now three different components. The left hand side has the TSS or total sum of squares which is the model's total variation. First on the right hand side is ESS or Explained Sum of

NOTES

Squares and next to it is RSS or the Residual Sum of Squares which depicts the variation which is unexplained.

A point to note: For RSS and ESS, the notation that is used can vary. So, it is best to be aware of what the two actually denote.

Now, the revealed identity can be depicted as follows:

$TSS = ESS + RSS$

and can also be rewritten as:

$$\frac{TSS}{TSS} = \frac{ESS}{TSS} + \frac{RSS}{TSS} = 1$$

So, it is possible to express the unexplained and explained variation, through dividing the total variation on both sides, as shares of the total variation.

Variation percent within the dependent variable connected to or explained by variation within the independent variable of the regression equation:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}, \quad 0 \le R^2 \le 1$$

Example 4.3: Consider that there is a simple linear regression model which estimated $R^2 = 0.65$. This denotes that from the total variation around mean of *Y*, 65 per cent can be explained with the variable *X* included in the model.

Within this simple model of regression, a good relationship exists between the coefficient of determination, OLS estimator of the slope and the coefficient measures of sample correlation coefficient. For this to become clear, let us explain the sum of squares as follows:

$$ESS = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(\left(b_0 + b_1 X_i\right) - \left(b_0 + b_1 \bar{X}\right)\right)^2 = \sum_{i=1}^{n}\left(b_1 X_i - b_1 \bar{X}\right)^2 = b_1^2 \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

With the help of this transformation, the coefficient of determination can be further expressed as:

$$R^2 = \frac{ESS}{TSS} = \frac{b_1^2 \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} = \left(b_1 \frac{S_X}{S_Y}\right)^2$$

In the above, *SX* and *SY* denote the sample standard deviation for *X* and *Y*, respectively. Following is the establishing of a relation between the OLS slope estimator and the correlation coefficient between *X* and *Y*.

$$b_1 = \frac{S_{XY}}{S_X^2} = \frac{S_{XY}}{S_X^2} \times \frac{S_Y}{S_Y} = \frac{S_{XY}}{S_X S_Y} \times \frac{S_Y}{S_X} = r\frac{S_Y}{S_X}$$

In the above, *SXY* denote the sample covariance between *X* and Y, and *r*; which is the sample correlation coefficient for *X* and for Y. Therefore, when we

NOTES

substitute the above two equations, we get the relation between the coefficient of determination and the sample correlation coefficient.

$$R^2 = \left(b_1 \frac{S_X}{S_Y}\right)^2 = \left(r \times \frac{S_X}{S_Y} \times \frac{S_Y}{S_X}\right)^2 = (r)^2$$

So, in the case of simple regression, absolute value of the sample correlation coefficient forms the square root of the coefficient of determination:

$$|r| = \sqrt{R^2}$$

It would imply that the larger is the correlation between X and Y, the larger will be the explained share of the model's variation (the smaller is the variation's unexplained share). Further, the less disperse will be the sample points from the regression line, the larger is the correlation and the coefficient of determination.

4.6.1 R^2 and the Significance of the OLS Estimators

If the variation in Y is increased while variation in X is unchanged, it will decrease the size directly of the coefficient of determination. Though it will not affect the importance of the regression model's parameter estimate.

Looking at the above equations makes it evident that variation increase in Y causes a reduction in the regression model's size of the coefficient of determination. Nevertheless, with increase in the variation in Y, the covariance between Y and X also increases and this raises the value of the parameter estimate. Therefore, it isn't evident that there will not be a change in the parameter's significance. With f-ratio creation, the following becomes evident:

$$t = \frac{b_1}{se(b_1)} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \Bigg/ \sqrt{\frac{S^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\underbrace{\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}}_{s} \times \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

In the above, S denoted the residual's **standard deviation**. Let us look at the outcome for the f-value when variation of Y is increased with a constant c.

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})(cY_i - c\bar{Y})}{\sqrt{\frac{\sum_{i=1}^{n}(cY_i - c\hat{Y}_i)^2}{n-2}} \times \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \frac{\sum_{i=1}^{n}c(X_t - \bar{X})(Y_t - \hat{Y})}{\sqrt{\frac{c^2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} \times \sqrt{\sum_{i=1}^{n}(X_t - \bar{X})^2}} = \frac{c}{c} \times \frac{b_1}{se(b_1)} = t$$

Therefore, raising the variation of Y with a constant c, does not in any way affect the f-value. So, the conclusion can be drawn that the coefficient of determination is no more than a measure of linear strength of the model.

NOTES

4.6.2 Adjusted R Square

The adjusted R^2 (also depicted as $\bar{R}^2$ and called 'R bar squared') is used to try to account for the phenomenon of the R^2 increasing spuriously and automatically on addition to the model of extra explanatory variables. This modification is caused by Theil's adjusted R^2 which adjusts to the number of explanatory terms in a model in relation with number of data points. There is an increase in the value of adjusted R^2 only in case the increase in R^2 (because of including a new explanatory variable) is greater than would have happened by chance. In case we introduce a set of explanatory variables that has a predetermined hierarchy of importance, one at a time into a regression, with the adjusted R^2 being computed at every introduction, the level at which adjusted R^2 reaches a maximum, and decreases afterward, would be the regression with the ideal combination of having the best fit without excess/ unnecessary terms. So, adjusted R^2 is defined as:

$$\bar{R}^2 = 1-\left(1-R^2\right)\frac{n-1}{n-p-1} = R^2-\left(1-R^2\right)\frac{p}{n-p-1}$$

In the above:

- p represents the number of explanatory variables of the model (excluding constant term)
- n represents sample size

We can also write adjusted R^2 as below:

$$\bar{R}^2 = 1-\frac{SS_{res}/df_e}{SS_{tot}/df_t}$$

In the above:

- df_t represents the degrees of freedom $n-1$ of the estimate of the population variance of the dependent variable
- df_e represents the degrees of freedom $n-p-1$ of the estimate of the underlying population error variance.

It is possible to see the fundamental principle of the adjusted R^2 statistic if ordinary R^2 is rewritten as follows:

$$R^2 = 1-\frac{VAR_{res}}{VAR_{tot}}$$

In the above:

$VAR_{res} = SS_{res}/n$ is the sample variances of the estimated residuals and $VAR_{tot} = SS_{tot}$ is the sample variances of the dependent variable. These can be looked upon as being biased estimates of the population variances of the errors and dependent variable.

NOTES

Below is how we can replace the estimates with statistically unbiased versions:

$VAR_{res} = SS_{res}/(n - p - 1)$

and

$VAR_{tot} = SS_{tot}/(n - 1)$

The interpretation for adjusted R^2 is not the same as for R^2. R^2 is a measure of fit while adjusted R^2 is a comparative measure of suitability of alternative nested sets of explanators. It is important to be careful while interpreting and reporting this statistic. Adjusted R^2 proves to be of great use at the stage of feature selection in building of a model.

Check Your Progress

10. Elaborate on the sample coefficient of determination?
11. Define the term adjust R square.

4.7 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. In the multiple regression model, there is extension of the simple (two-variable) regression model to further take into account possibility of additional explanatory factors which can systematically affect the dependent variable.
2. We carry out the minimization through differentiating the scalar S for all $b's$ turn wise, and making the first order condition that results a zero. So we get, (k+1) simultaneous equations in (k+1) unknowns, called the least squares normal equations.
3. Positive square root of s^2 is called standard error of regression (SER). The units of SER are same as of the dependent variable. It is the numerator of the estimated standard errors of the OLS coefficients. Oftentimes, SER's magnitude is compared with the dependent variable's mean to determine the ability of the regression for explaining the data.
4. The aim of LSM method under SLR is to estimate the parameters of the model, based on the observed pairs of values and applying a certain criterium function (the observed pairs of values are constituted by selected values of the auxiliary variable and by the corresponding observed values of the response variable), that is:

   Observed values- x_i and y_i for each pair i, where i=1,2,...,i,...n

   Values to be estimated A and B and $(Y_1, Y_2, ..., Y_i, ..., Y_n)$ for the n observed pairs of values.

NOTES

5. In the least squares method the estimators are the values of A and B which minimize the object function. Thus, one has to calculate the derivatives $\partial\Phi/\partial A$ e $\partial\Phi/\partial B$, equate them to zero and solve the system of equations in A and B.
6. The aim of the Leat Squares method under the MLR is to estimate the parameters of the model**,** based on the observed **n** sets of values and by applying a certain criterium function (the observed sets of values are constituted by selected values of the auxiliary variable and by the corresponding observed values of the response variable), that is:

   Observed values $x_{1,i}\ x_{2,i,...},\ x_{j,i,..},\ x_{k,i}$ and y_i for each set i, where i=1,2,...,i,...n

   Values to be estimated $B_1,B_2,...,B_j,...,B_k$ et $(Y_1,Y_2,..., Y_i,..., Y_n)$
7. The Ford-Walford and Gulland and Holt expressions, are already in their linear form, allowing the estimation of K and L_∞ with methods of simple linear regression on observed L_i and T_i. The Gulland and Holt expression allows the estimation of K and L_∞ even when the intervals of time T_i are not constant. In this case, it is convenient to re-write the expression as:

   $$\Delta L / T_i = K.L_\infty - K.\overline{L}$$
8. Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable. Multiple Liner Regression **(**MLR) is used extensively in econometrics and financial inference.
9. **Variance of $\hat{Y}$**

   The variance of $\hat{Y}$ **is**

   $$\begin{aligned} V(\hat{y}) &= V(Xb) \\ &= XV(b)X' \\ &= \sigma^2 X(X'X)^{-1}X' \\ &= \sigma^2 H. \end{aligned}$$
10. The coefficient of determination, R^2, is a widely used measure of the goodness of fit of a regression line. It measures the extent or strength of the association that exists between the two variables, X and Y. Since this value is based on the sample data points, it is known as the 'Sample Coefficient of determination'. It's value ranges from 0 (poor fit) to 1 (good fit) or (perfect fit).
11. The adjusted R^2 (also depicted as $\overline{R}^2$ and called 'R bar squared') is used to try to account for the phenomenon of the R^2 increasing spuriously and automatically on addition to the model of extra explanatory variables. This modification is caused by Theil's adjusted R^2 which adjusts to the number of explanatory terms in a model in relation with number of data points.

4.8 SUMMARY

NOTES

- In the multiple regression model, there is extension of the simple (two-variable) regression model to further take into account possibility of additional explanatory factors which can systematically affect the dependent variable.
- It is possible to define the same three sums of squares (SST, SSE and SSR) even in multiple regression. R^2 is the ratio of the explained sum of squares (SSE) to the total sum of squares (SST). The correlation is now not simple though it even now possesses the interpretation of a squared simple correlation coefficient, which is the correlation between y and $\hat{y}$, $r_{\hat{y}y}$.
- In case the population intercept β_0 is not the same as zero, there will be bias in the slope coefficients that is computed via a regression through the origin. So, oftentimes, there is the inclusion of an intercept, and the data is allowed to decide if it should be zero.
- It is important to understand that this particular proposition is not an assumption about the population model but rather an implication of the sample data being used. Further, this will be valid only in the case of linear relations among the explanatory variables.
- Generally, each OLS coefficient of an underspecified model will show bias in such a situation unless the variable that has been omitted is uncorrelated with every included regressor. Generally, as a rule asymmetric nature of specification error can be taken away.
- Economic, Social and Scientific variables are often times connected by linear associations with the assumption that the model is linear in parameters. In the field of econometrics and statistical modeling, regression analysis is a conceptual process for ascertaining the functional relationship that exists among variables.
- The least squares method is presented under the forms of Simple linear Regression, multiple linear model and non-linear models (method of Gauss-Newton).
- To calculate the least squares estimators it will suffice to put the derivative $d\Phi/dB$ of Φ in order to vector B, equal to zero. $d\Phi/dB$ is a vector with components $\partial\Phi/\partial B_1$, $\partial\Phi/\partial B_2$,..., $\partial\Phi/\partial B_k$.
- In statistical analysis it is convenient to write the estimators and the sums of the squares using idempotent matrices. Then the idempotent matrices L, (I - L) and (I - M) with $L_{(n,n)} = X\,(X^{T.}\,X)^{-1.}\,X^T$, I = unity matrix and $M_{(n,n)}$ = $mean_{(n,1)}$ matrix = 1/n [1] where [1] is a matrix with all its elements equal to one, are used.

- Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable. Multiple Liner Regression (MLR) is used extensively in econometrics and financial inference.
- The variance of b can be obtained as the sum of variances of all b_1, b_2,,..., b_k which is the trace of covariance matrix of b.
- The coefficient of determination, R^2, is a widely used measure of the goodness of fit of a regression line. It measures the extent or strength of the association that exists between the two variables, X and Y. Since this value is based on the sample data points, it is known as the 'Sample Coefficient of determination'. It's value ranges from 0 (poor fit) to 1 (good fit) or (perfect fit).
- R^2 is defined as the proportion of the total variation, i.e., 'explained' and 'unexplained'.

  i.e; $\Sigma(Y - \bar{Y})^2 = \Sigma(\hat{Y} - \bar{Y})^2 + \Sigma(Y - \hat{Y})^2$

  Thus, $$R^2 = \frac{\text{total variation} - \text{unexplained variation}}{\text{total variation}}$$

  $$= 1 - \frac{\Sigma(Y - \hat{Y})^2}{\Sigma(Y - \bar{Y})^2}$$

- The adjustment factor being $(n - 1)/(n - 2) > 1$, R^2 will always be less than or equal to $\bar{R}^2$ as the sample size increases, the ratio $\frac{(n-1)}{(n-2)} \to 1$ and thus, the difference between R^2 and $\bar{R}^2$ will be reduced.
- For a simple regression mode, it is possible to explain the variation of one variable with the help of another due to the fact that they have a correlation. In case they were not correlated, no explanatory power would be there in the X variable. During regression analysis, there is a clear relation between the correlation coefficient and the coefficient of determination, though there is a slight difference in their interpretation.
- No information is provided by the correlation coefficient regarding two variables' causal relationship. Let us look at the correlation coefficient in a context of the regression model and know in what circumstances it will be right to interpret the correlation coefficient as a measure of strength of a causal relationship.
- In **coefficient of determination**, the average deviation from mean is split into two parts: explained and an unexplained part. So, it is natural to start the derivation of the measure from the deviation from the mean expression, following which, the predicted value obtained from the regression model are introduced.

- The deviation from the mean value of Y is explained by employing the regression model. So, we will denote the difference between the expected value (y) and the mean value *(y)* as the mean difference's explained part. The part that remains is the unexplained part.
- After the manipulations have been made, there are now three different components. The left hand side has the TSS or total sum of squares which is the model's total variation. First on the right hand side is ESS or Explained Sum of Squares and next to it is RSS or the Residual Sum of Squares which depicts the variation which is unexplained.
- If the variation in Y is increased while variation in X is unchanged, it will decrease the size directly of the coefficient of determination. Though it will not affect the importance of the regression model's parameter estimate.
- The interpretation for adjusted R^2 is not the same as for R^2. R^2 is a measure of fit while adjusted R^2 is a comparative measure of suitability of alternative nested sets of explanators. It is important to be careful while interpreting and reporting this statistic. Adjusted R^2 proves to be of great use at the stage of feature selection in building of a model.

4.9 KEY WORDS

- **Multiple regression model:** Multiple regression model, there is extension of the simple (two-variable) regression model to further take into account possibility of additional explanatory factors which can systematically affect the dependent variable.
- **Least squares method:** The least squares method is presented under the forms of Simple linear Regression, multiple linear model and non-linear models (method of Gauss-Newton).
- **Variance of b** The variance of b can be obtained as the sum of variances of all b_1, $b_{2,}$,..., b_k which is the trace of covariance matrix of b.
- **Coefficient of determination:** The coefficient of determination, R^2, is a widely used measure of the goodness of fit of a regression line. It measures the extent or strength of the association that exists between the two variables, X and Y. Since this value is based on the sample data points, it is known as the 'Sample Coefficient of Determination'.

4.10 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Define the multiple regression model.

2. Elaborate on the mechanics and interpretation of OLS.
3. Interpret the fitted value and residuals.
4. Give the assumptions of multiple regression model.
5. What is simple linear regression model?
6. Elaborate on the ordinary least squares.
7. What do you understand by estimation of parameter based on OLS?
8. Elaborate on the MLR of OLS model.
9. Interpret the estimation method.
10. What is residual sum of squares?
11. Explain the non-linear model.
12. Give the principle of OLS.
13. Explain about the properties of OLSE.
14. What is goodness of fit?
15. Define the term $R^{2.}$
16. Comprehend the adjust R square.

Long-Answer Questions

1. Briefly explain about the multiple linear regression model with the appropriate examples.
2. Discuss in detail the fitted values and residuals and their properties.
3. Analyse the different types of estimation of parameters based on OLS model.
4. Explain briefly about the method of Gauss-Newton giving its advantages.
5. Discuss in detail about the properties of MLR based on OLS with the help of examples.
6. Briefly explain about the sample coefficient determination of goodness of fit with the help of examples.
7. Explain in detail about the goodness of fit with various examples.
8. Distinguish between R^2 and adjust R square with appropriate examples.

4.11 FURTHER READINGS

Johnston, J. and John DiNARDO. 1997. *Econometric Methods*, Fourth Edition. New Delhi: Tata McGraw-Hill.

Koutsoyiannis, A.1977. *Theory of Econometrics,* Second Edition. London: The Macmillan Press Ltd.

NOTES

Özdemir, Durmuº. 2016. *Applied Statistics for Economics and Business,* Second Edition. Izmir (Turkey): Springer.

Maddala, G. S. 1992. *Introduction to Econometrics,* Second Edition. New York: Macmillan Publishing Company.

Pindyck, R. S and D. L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts,* Fourth Edition. New York: McGraw Hill.

Goldberger, A. S. 1998. *Introductory Econometrics*. Cambridge: Harvard University Press.

Levine, David M., Timothy C. Krehbiei, Mark L. Berenson and P. K. Viswanathan. 2009. *Business Statistics*, Fifth Edition. New Delhi: Pearson Education.

Webster, Allen L. 1998. *Applied Statistics for Business and Economics,* Third Edition. New Delhi: Tata McGraw-Hill.

NOTES

BLOCK - III

ECONOMETRIC ANALYSIS

UNIT 5 VIOLATIONS OF CLASSICAL ASSUMPTIONS

Structure

5.0 INTRODUCTION

One of the assumptions of the classical linear regression model is that there is no heteroscedasticity. Breaking this assumption means that the Gauss–Markov theorem does not apply, meaning that Ordinary Least Squares (OLS) estimators are not the Best Linear Unbiased Estimators (BLUE) and their variance is not the lowest of all other unbiased estimators. Heteroscedasticity does not because ordinary least squares coefficient estimates to be biased, although it can because ordinary least squares estimates of the variance (and, thus, standard errors) of the coefficients to be biased, possibly above or below the true of population variance.

Thus, regression analysis using heteroscedastic data will still provide an unbiased estimate for the relationship between the predictor variable and the outcome, but standard errors and therefore inferences obtained from data analysis are suspect. Biased standard errors lead to biased inference, so results of hypothesis tests are possibly wrong. For example, if OLS is performed on a heteroscedastic data set, yielding biased standard error estimation, a researcher might fail to reject a null hypothesis at a given significance level, when that null hypothesis was actually uncharacteristic of the actual population (making a type II error).

The existence of heteroscedasticity is a major concern in regression analysis and the analysis of variance, as it invalidates statistical tests of significance that

assume that the modelling errors all have the same variance. While the ordinary least squares estimator is still unbiased in the presence of heteroscedasticity, it is inefficient and generalized least squares should be used instead.

NOTES

Multicollinearity refers to a situation in which more than two explanatory variables in a multiple regression model are highly linearly related. We have perfect multicollinearity if, for example as in the equation above, the correlation between two independent variables is equal to 1 or –1. In practice, we rarely face perfect multicollinearity in a data set. More commonly, the issue of multicollinearity arises when there is an approximate linear relationship among two or more independent variables.

In this unit, you will study about the violations of classical assumptions, consequences, detection and remedies multicollinearity, heteroscedasticity, and serial correlations.

5.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the violations of classical assumptions
- Elaborate on the consequences
- Explain the detection and remedies multicollinearity
- Define the heteroscedasticity
- Analyse the serial correlations

5.2 VIOLATIONS OF CLASSICAL ASSUMPTIONS CONSEQUENCES

There are several different frameworks in which the linear regression model can be cast in order to make the OLS technique applicable. Each of these settings produces the same formulas and same results. The only difference is the interpretation and the assumptions which have to be imposed in order for the method to give meaningful results. The choice of the applicable framework depends mostly on the nature of data in hand, and on the inference task which has to be performed.

One of the lines of difference in interpretation is whether to treat the regressors as random variables, or as predefined constants. In the first case (random design) the regressors xi are random and sampled together with the yi's from some population, as in an observational study. This approach allows for more natural study of the asymptotic properties of the estimators. In the other interpretation (fixed design), the regressors X are treated as known constants set by a design, and y is sampled conditionally on the values of X as in an experiment. For practical purposes, this distinction is often unimportant, since estimation and inference is carried out

while conditioning on X. All results stated in this article are within the random design framework.

The classical model focuses on the "Finite Sample" estimation and inference, meaning that the number of observations n is fixed. This contrasts with the other approaches, which study the asymptotic behaviour of OLS, and in which the number of observations is allowed to grow to infinity. The errors in the regression should have conditional mean zero: $\mathrm{E}[\varepsilon \mid X] = 0.$

The immediate consequence of the exogeneity assumption is that the errors have mean zero: $E[\varepsilon] = 0$, and that the regressors are uncorrelated with the errors: $E[X^T\varepsilon] = 0$.

The exogeneity assumption is critical for the OLS theory. If it holds then the regressor variables are called exogenous. If it doesn't, then those regressors that are correlated with the error term are called endogenous, and then the OLS estimates become invalid. In such case the method of instrumental variables may be used to carry out inference.

It is sometimes additionally assumed that the errors have normal distribution conditional on the regressors: $\varepsilon \mid X \sim \mathcal{N}(0, \sigma^2 I_n).$

This assumption is not needed for the validity of the OLS method, although certain additional finite-sample properties can be established in case when it does (especially in the area of hypotheses testing). Also when the errors are normal, the OLS estimator is equivalent to the Maximum Likelihood Estimator (MLE), and therefore it is asymptotically efficient in the class of all regular estimators. Importantly, the normality assumption applies only to the error terms; contrary to a popular misconception, the response (dependent) variable is not required to be normally distributed.

One consequence of a high degree of multicollinearity is that, even if the matrix $X^\top X$ is invertible, a computer algorithm may be unsuccessful in obtaining an approximate inverse, and if it does obtain one it may be numerically inaccurate. But even in the presence of an accurate matrix, the following consequences arise.

In the presence of multicollinearity, the estimate of one variable's impact on the dependent variable *Y* while controlling for the others tends to be less precise than if predictors were uncorrelated with one another. The usual interpretation of a regression coefficient is that it provides an estimate of the effect of a one unit change in an independent variable, X_1 holding the other variables constant. If X_1 is highly correlated with another independent variable, X_2 in the given data set, then we have a set of observations for which X_1 and X_2 have a particular linear stochastic relationship. We don't have a set of observations for which all changes in X_1 are independent of changes in X_2 so we have an imprecise estimate of the effect of independent changes in X_1.

In some sense, the collinear variables contain the same information about the dependent variable. If nominally "Different" measures actually quantify the same phenomenon then they are redundant. Alternatively, if the variables are accorded different names and perhaps employ different numeric measurement scales but are highly correlated with each other, then they suffer from redundancy. One of the features of multicollinearity is that the standard errors of the affected coefficients tend to be large. In that case, the test of the hypothesis that the coefficient is equal to zero may lead to a failure to reject a false null hypothesis of no effect of the explanator, a type II error. Another issue with multicollinearity is that small changes to the input data can lead to large changes in the model, even resulting in changes of sign of parameter estimates.

A principal danger of such data redundancy is that of over fitting in regression analysis models. The best regression models are those in which the predictor variables each correlate highly with the dependent (outcome) variable but correlate at most only minimally with each other. Such a model is often called "Low Noise" and will be statistically robust (that is, it will predict reliably across numerous samples of variable sets drawn from the same statistical population).

So long as the underlying specification is correct, multicollinearity does not actually bias results; it just produces large standard errors in the related independent variables. More importantly, the usual use of regression is to take coefficients from the model and then apply them to other data. Since multicollinearity causes imprecise estimates of coefficient values, the resulting out-of-sample predictions will also be imprecise. And if the pattern of multicollinearity in the new data differs from that in the data that was fitted, such extrapolation may introduce large errors in the predictions.

5.3 DETECTION AND REMEDIES MULTICOLLINEARITY

In the field of statistics, multicollinearity refers to the phenomenon where, in a multiple regression model, two or a greater number of predictor variables are highly correlated. This implies that one can be linearly predicted from the others with accuracy of a high degree. In a situation of this kind, the multiple regression's coefficient estimates could erratically change as a response to small changes in the data's model. The reliability or predictive power is not decreased by multicollinearity as far as the model as a whole is concerned, at least within the sample data set; it just has an effect on the calculations pertaining to individual predictors. In other words, a multiple regression model with correlated predictors can show how well the total set of predictors predict the outcome variable, yet, it might not provide results that are valid for individual predictor, or about which predictors are redundant with respect to others.

When we consider perfect multicollinearity, the predictor matrix is singular due to which it cannot get inverted. In such circumstances, for a general linear model $y = X\beta + \in$, the ordinary least-squares estimator $\hat{\beta}_{OLS} = (X'X)^{-1}X'y$ is non-existent.

It must be noted that in case of statements of the assumptions that underlie regression analyses, the use of 'no multicollinearity' is sometimes meant to depict the absence of perfect multicollinearity, which is an exact (non-stochastic) linear relation among the regressors.

Nature and Estimation in its Presence

A linear association between two explanatory variables is known as collinearity. We will consider two variables to be perfectly collinear when an exact linear relationship exists between the two. To take an example, X_1 and X_2, will be considered to be perfectly collinear when there are parameters λ_0 and λ_1 so that, for every observation i, there is:

$$X_{2i} = \lambda_0 + \lambda_1 X_{1i}$$

A situation where two or more explanatory variables in a multiple regression model are highly linearly related, it is referred to as multicollinearity. There is perfect multicollinearity if, for example, as depicted by the above equation, between two independent variables the correlation equals 1 or –1.

In actual practice, a data set rarely depicts perfect multicollinearity. It is more commonly seen that the issue of multicollinearity arises in the case where between two or more independent variables, there is an approximate linear relationship.

In mathematical terms, we will consider a set of variables to be perfectly multicollinear in the case where, among some of the variables, there exists one or more exact linear relationships. To take an example, there might be:

$$\lambda_0 + \lambda_1 X_{1i} + \lambda_2 X_{2i} + \cdots + \lambda_k X_{ki} = 0.$$

This holds for all observations i, where λ_j are constants and X_{ji} is the i^{th} observation on the j^{th} explanatory variable. It is possible to explore one issue created by multicollinearity while studying the process used to attain estimates of the multiple regression equation parameters

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i.$$

In ordinary least squares estimates, there is a need to invert the matrix

$$X^T X$$

Where:

$$X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1N} & \cdots & X_{kN} \end{bmatrix}.$$

NOTES

In case between independent variables, an exact linear relationship (perfect multicollinearity) exists, the rank of X (also of X^TX) will be less than k+1, and the X^TX matrix shall not be invertible.

While using raw data sets, it is common to face perfect multicollinearity. Raw data sets oftentimes have redundant information. After identification and removal of redundancies has been done, there will be nearly multicollinear variables still there because of the correlations that are inherent to the system being studied. In a situation like this, rather than the equation that has been present above being applicable, the equation can be modified with an error term v_i:

$$\lambda_0 + \lambda_1 X_{1i} + \lambda_2 X_{2i} + \cdots + \lambda_k X_{ki} + v_i = 0.$$

For a case like this, there will not exist any exact linear relationship among the variables, though the variables X_j are nearly perfectly multicollinear when v_i's variance is small in case of some set of values of λ's. In such situations, there is an inverse for the matrix X^TX, though it is ill-conditioned so that a given computer algorithm may or may not be able to compute an approximate inverse. In case it does, the resulting computed inverse may be highly sensitive to slight variations in the data (because the rounding error's effect gets magnified), making it possibly greatly inaccurate.

Effects of multicollinearity

An effect of high degree of multicollinearity is that, despite matrix X^TX being invertible, the computer algorithm might not be successful in getting an approximate inverse, and in case it is obtained by the algorithm, it might lack numerical inaccuracy. Yet, even if the X^TX matrix is accurate, there is the possibility of several effects. Let us look at these briefly.

When there is multicollinearity, estimating what effect a variable will have on a dependent variable Y while controlling for others, is not as precise than it would be if the predictors were not correlated with each other. Generally, the interpretation of a regression coefficient provides an estimate of the effect of a one unit change in an independent variable,X_1, while holding the other variables constant. In case of X_1 being highly correlated with another independent variable, X_2, in the given data set, there will be a set of observations for which $X_1$and X_2 will possess a specific linear stochastic relationship. If the observations do not suggest such that every change in X_1 is not dependent on changes in X_2, there will be an imprecise estimate of the effect of independent changes in X_1.

In a way, the same information is present in the collinear variables about the dependent variable. If 'different' measures actually quantify the same phenomenon, then they are redundant. On the other hand, in case different names are given to the variables and possibly use differing scales for numeric measurement, yet remain highly correlated with each other, then they still have redundancy.

In multicollinearity, the affected coefficients' standard errors are large. So, in such a situation, the testing for the hypothesis of the coefficient being zero could

cause a failure to reject a false null hypothesis of no effect of the explanatory. This will be a type II error.

Moreover, with multicollinearity the issue is that even little changes made to the input data are capable of causing large changes in the model. These large changes in the model can even lead to changes of sign of parameter estimates.

One key problem with this kind of data redundancy is what is known as overfitting in regression analysis models. In regression models, the best one are the ones where every predictor variable correlates highly with the dependent (outcome) variable. With each other, at most, they correlate only minimally. A model of this kind is generally referred to as 'low noise' and is mostly robust statistically. It is therefore capable of reliably predicting across numerous samples of variable sets taken from the same statistical population.

Till the underlying specification is correct, results are not biased by multicollinearity, but rather it causes large standard errors in the related independent variables. Furthermore, regression is generally used to take coefficients from the model and apply them to other data. In case there is a difference in the pattern of multicollinearity in the fitted and the new data, large errors might get introduced in the predictions.

Detecting Multicollinearity

The following are the indications of the possible presence of multicollinearity within the model:

1. **Predictor Variable:** On deletion or addition of a predictor variable, there is a big change in the estimated regression coefficients.
2. **Multiple Regression:** While affected variables in the multiple regression have regression coefficients that are insignificant, there is a rejection of the joint hypothesis that those coefficients are all zero (by employing the *F*-test).
3. **Multivariable Regression:** For a multivariable regression, if there is an insignificant coefficient of a particular explanator, and there is a simple linear regression of the explained variable on this explanatory variable which shows its coefficient to be significantly different from zero, this shows multicollinearity in the multivariable regression.
4. **VIF:** It has been suggested that there be a formal detection-tolerance or the Variance Inflation Factor (VIF) for multicollinearity:

   $$\text{tolerance} = 1 - R_j^2, \quad \text{VIF} = \frac{1}{\text{tolerance}},$$

   In the above, R_j^2 represents coefficient of determination of a regression of explanatory *j* on all the other explanators. Tolerance below 0.20 or 0.10 and/or a VIF of 5 or 10 or more, points to a problem of multicollinearity.

5. **The condition number test:** The standard measure of ill-conditioning in a matrix is the condition index. This points to the matrix's inversion being unstable numerically with finite-precision numbers (standard computer floats and doubles). It shows that there is potential sensitivity of the computed inverse towards small changes in the original matrix. It is possible to compute the condition number through locating the square root of the maximum eigenvalue divided by the minimum eigenvalue. When a Condition Number is more then 30, there might be significant multicollinearity in the regression; there is the presence of multicollinearity, in case, additionally 2 or greater number of variables associated with the high condition number have high proportions of variance explained. This method has one advantage: it can identify which variables are the cause of the problem.
6. **Using the Farrar–Glauber test:** In case the variables are orthogonal, multicollinearity does not exist; in case the variables are not orthogonal, it indicates the presence of multicollinearity. It has been opined by C. Robert Wichers that the 'Farrar–Glauber partial correlation test is ineffective in that a given partial correlation may be compatible with different multicollinearity patterns.' There has also been a lot of criticism of the Farrar–Glauber test.
7. **Perturbing the data:** It is possible to detect multicollinearity with introducing random noise to the data and performing the regression several times to check the amount of change in the coefficients.
8. If a correlation matrix is constructed among the explanatory variables, it will provide pointers to the possibility of any given couplet of right-hand-side variables causing problems of multicollinearity. Correlation values of a minimum .4 can be considered as pointing to a problem of multicollinearity. Nevertheless, such a procedure is extremely problematic and not recommended for use. Intuitively, collinearity is a multivariate phenomenon, while on the other hand, correlation will describe a relationship that is bivariate.

Remedies for Multicollinearity

1. Stay away from the dummy variable trap; including a dummy variable for every category and including a constant term in the regression will ensure perfect multicollinearity.
2. Check what will happen when use is made of independent subsets of the data for estimation and the estimates are applied to the entire data set. In theory, what should be obtained is a somewhat higher variance from the smaller data sets used for estimation, but it is essential that the expectation of the coefficient values is same. While the observed coefficient values will vary, check by how much they vary.

3. Even with the presence of multicollinearity, maintain the model as it is. Multicollinearity will not have an effect in the efficacy of extrapolating the fitted model to new data if the predictor variables follow the same pattern of multicollinearity in the new data as they did in the data that the regression model was based upon.
4. One option could be to drop one of the variables. It is possible to drop an explanatory variable to create a model with significant coefficients. But this will lead to information loss as a variable has been dropped. If a relevant variable has been dropped, it will lead to biased coefficient estimates for the remaining explanatory variables which were correlated with the variable that was omitted.
5. A perfect solution would be to gather more data, wherever possible. With more data, it is possible to have more precise parameter estimates, and lower standard errors.
6. Another remedy could be the predictor variables' mean-centering. Collecting polynomial terms could create a certain amount of multicollinearity in case the specific variable has a limited range. With mean-centering, it is possible to remove this special type of multicollinearity. Nevertheless, overall, this will not cause any effect. It can help to remove such problems that are caused by computational steps, such as rounding, if use is not made of a well-designed computer program.
7. Independent variables must be standardized. It could decrease the false flagging of a condition index above 30.
8. It is possible that employing Shapley value, a tool used in game theory, can account for the effects of multicollinearity. With the Shapley value, a value is assigned for every predictor and every possible combinations of importance is assessed.
9. Use can be made of partial least squares regression or principal component regression or ridge regression.
10. In case of the correlated explanators being different lagged values of the same underlying explanator, use can be made of a distributed lag, which will impose a general structure on the relative values of the coefficients to be estimated.

    It is important to remember that orthogonalizing the explanatory variables does not help in offsetting the effects of multicollinearity.

Example 5.1: Consider the following data pertaining to consumption, income and wealth. All the data is in INR.

NOTES

$Y_,$	$X_{2'}$	$X_{3'}$
70	80	810
65	100	1009
90	120	1273
95	140	1425
110	160	1633
115	180	1876
120	200	2052
140	220	2201
155	240	2435
150	260	2686

The formula applied for obtaining the standard error is:

$$se\left(\hat{\beta}_2 + \hat{\beta}_3\right) = \sqrt{\text{var}\left(\hat{\beta}_2\right) + \text{var}\left(\hat{\beta}_3\right) + 2\ \text{cav}\left(\hat{\beta}_2, \hat{\beta}_3\right)}$$

In case we go with the assumption that there is a linear relationship between expenditure on the one hand and wealth and income on the other, the regression for the above date will be as follows:

$$\begin{aligned} \hat{Y}_i &= 24.7747 + 0.9415\ X_{2i} - 0.0424\ X_{3i} \\ &\quad (6.7525) \quad (0.8229) \qquad (0.0807) \\ t &= (3.6690) \quad (1.1442) \qquad (-0.5261) \\ &\quad R^2 = 0.9635 \quad \bar{R}^2 = 0.9531 \quad \text{df} = 7 \end{aligned}$$

From the above regression it can be seen that that approximately 96 per cent of consumption expenditure variation is dependent on wealth and income. Despite this, neither of the slope coefficients is individually significant in terms of statistics. Furthermore, along with being statistically insignificant, the wealth variable even has the wrong sign.

A Priori, it would be expected that wealth and consumption have a positive relationship.

While individually $\hat{\beta}_2$ and $\hat{\beta}_3$ are statistically insignificant, on testing the hypothesis of $\beta_2 = \beta_3 = 0$ simultaneously, we could reject the hypothesis depicted in the table given below:

ANOVA Table Depicting the Example for Consumption-Income-Wealth

Source of Variation	SS	df	MSS
Due to regression	8,565.5541	2	4,282.7770
Due to residual	324.4459	7	46.3494

The unusual assumption provides:

$$F = \frac{4282.7770}{46.3494} = 92.4019$$

The F value is of great significance.

Let us check this result geometrically. The intervals depict that each individually includes the value zero. Thus, the hypothesis that the two partial slopes are zero can be individually accepted. But then the joint confidence interval is established for testing the hypothesis $\beta_2 = \beta_3 = 0$, is not possible to accept the hypothesis as the joint confidence interval, in ellipse, has excluded the origin. Under high collinearity, tests will not be reliable on individual regressors, and here the overall F test will bring out if there is a relation of Y with the different regressors.

The example clearly depicts what happens with multicollinearity. It is important to note that the F test is significant despite the fact that the t values of both X_2 and X_3 are insignificant individually. This implies a high co-relation between the two variables making it impossible to isolate the impact that income and wealth will have on consumption. On regressing X_3 on X_2 it will produce the following, showing that X_3 on X_2 have perfect collinearity.

$$\hat{X}_{3i} = 7.5454 + 10.1909\, X_{2i}$$
$$(29.4758) \quad (0.1643)$$
$$t = (0.2560) \quad (62.0405)$$
$$R^2 = 0.9979$$

On regressing Y on X_2, we get the following which shows that the income variable is now significant:

$$\hat{Y}_i = 24.4545 + 0.5091\, X_{2i}$$
$$(6.4138) \quad (0.0357)$$
$$t = (3.8128) \quad (14.2432)$$
$$R^2 = 0.9621$$

On regressing Y on X_3, we get the following:

$$\hat{Y}_i = 24.411 + 0.0498\, X_{3i}$$
$$(6.874) \quad (0.0037)$$
$$t = (3.551) \quad (13.29)$$
$$R^2 = 0.9567$$

This shows that there is a significant impact on consumption expenditure of wealth. The above results show that to get out of extreme collinearity. One should just drop the collinear variable.

NOTES

5.4 HETEROSCEDASTICITY

The word 'heteroscedasticity' comes from the Greek, and quite literally means data with a different (hetero) dispersion (skedasis). In simple terms heteroscedasticity is any set of data that isn't homoscedastic. More technically, it refers to data with unequal variability (scalter) across a set of second, prefictor variables.

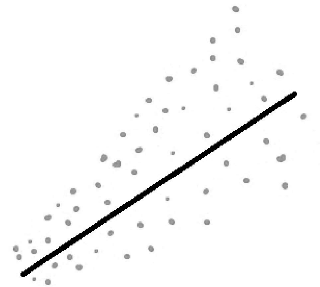


Fig. 5.1 *Heteroscedastic Data Tends to follow a Cone Shape on a Scalter Graph*

Nature and Estimation in its Presence

One of the important asumptions of the classical linear regression model is that the variance of each disturbance term μ_i, conditional on the chosen values of the explanatory variables, is some constant number equal to σ^2. This is the assumption of homoscedasticity, or equal (homo) spread (scedasticity), that is, equal variance.

Symbolically,

$$\in (\mu_i^2) = \sigma^2 \qquad i = 1, 2,...,N \qquad ...(5.1)$$

Figure 5.2 shows the conditional variance of Y_i (which is equal to that of μ_i), conditional upon the given X_i, remains the same regardless of the values taken by the variable X.

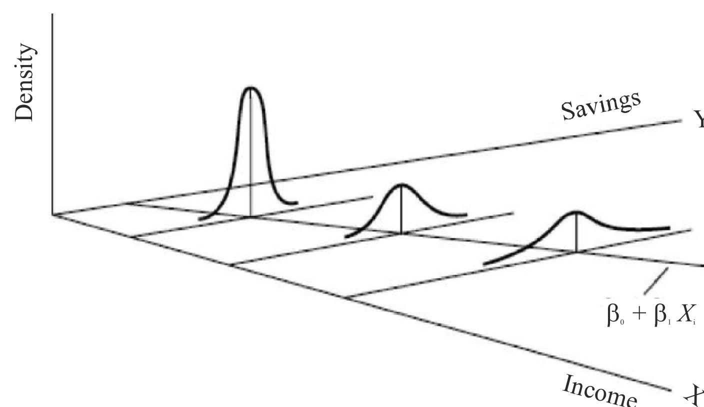


Fig. 5.2 *Homoscedastic Disturbances*

In contrast, consider Figure 5.3, which shows that the conditional variance of Y_i increases as X increases. Here, the variances of Y_i are not the same. Hence three is heteroscedasticity. Symbolically,

$$\in (\mu_i^2) = \sigma_i^2 \qquad ...(5.2)$$

Notice the subscript on σ^2, which reminds us that the conditional variances of μ_i (= conditional variance of Y_i) are no longer constant.

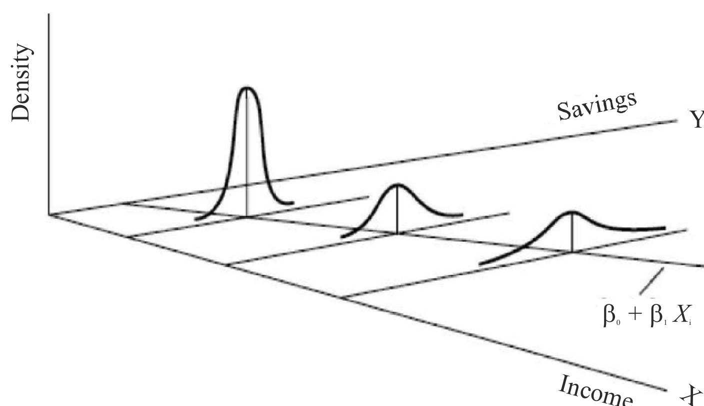


Fig. 5.3 *Heteroscedastic Distubances*

To make the difference between homoscedasticity and heteroscedasticity clear, assume that in the two-variable model $Y_i = \beta_0 + \beta_1 X_i + \mu_i$, Y represents savings and X represents income. Figure 5.2 and 5.3 show that as income increases, savings on the average also increase. But in Figure 5.2 the variance of savings remain the same at all levels of income, whereas in Figure 5.3 it increases with income. It seems that in Figure 5.3, the higher-income families on the average save more than the lower-income families, but there is also more variability in their savings.

There are several reasons why the variances of μ_i may be variable, some of which are as follows:

1. Following the error-learning models, as people learn, their errors of behaviour become smaller over time. In this case, σ_i^2 is expected to decreases. As an example, consider Figure 5.4, which relates the number of typing error made in a given time period on a test to the hours put in typie practice. As Figure 5.4 shows, as the number of hours of typing practice increases, the average number of typing errors as well as their variances decreases.

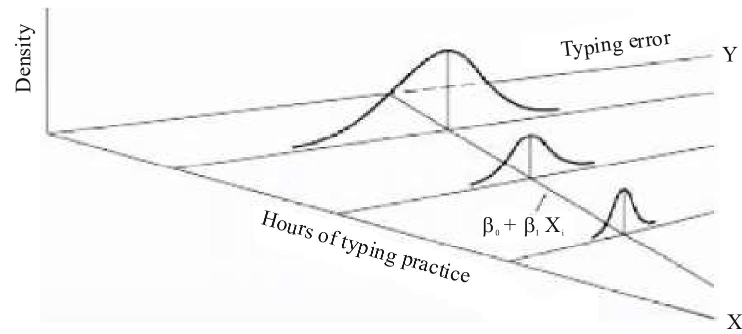


Fig. 5.4 *Illustration of Heteroscedasticity*

2. As incomes grow, people have more discretionary income and hence more scope for choice about the disposition of their income. Hence σ_i^2 is lkely to increase with income. Thus in the regression of savings on income one is likely to find σ_i^2 increasing with income (as in Figure 5.3) because people have more choices about their savings behaviour. Similarly, companies with larger profits are generally expected to show greater variability in their dividend policies than companies with lower profits. Also, growth-oriented companies are likely to show more variability in their dividend pay out ratio than established companies.
3. As data collecting techniques improve, σ_i^2 is likely to decrease. Thus banks which have sophisticated data processing equipment are likely to commit fewer errors in the monthly quarterly statements of their customers than banks without such facilities.

It should be noted that the problem of heteroscedasticity is likely to be more common in cross-sectional than time-series data. In cross-sectional data, one usually deals with members of a population at a given point in time, such as individual consumers or their families, firms, industries, or geographical subdivision, such as state, country, city, etc. Morever, these members may be of different sizes, such as small, medium, or large firms low, medium, or high income. In time-series data, on the other hand, the variables tend to be of similar order of magnitude because the data are generally collected for the same entity over a period of time.

As an illustration of heteroscedasticity likely to be encountered in cross-sectional analysis, consider Table 5.1. This table gives data on compensation per employee in to non-durable goods manufacturing industries classified by the employment size of the firm. Also given in the table are average productivity figures for the nine employment classes.

NOTES

Table 5.1 *Compensation Per Employee (in $) in Non-durable Manufacturing Industries According to Employment Size of Establishment*

	Employment size (Average number of employees)								
Industry	**1-4**	**5-9**	**10-19**	**20-49**	**50-99**	**100-249**	**250-499**	**500-999**	**1000-2499**
Food and Kindred Products	2994	3295	3565	3907	4189	4486	4676	4968	5342
Tobacco Products	1721	2057	3336	3320	2980	2848	3072	2969	3822
Textile mill products	3600	3657	3674	3437	3340	3334	3225	3163	3168
Apparel and related products	3494	3787	3533	3215	3030	2834	2750	2967	3453
Paper and allied products	3498	3847	3913	4135	4445	4885	5132	5342	5326
Printing and publishing	3611	4206	4695	5083	5301	5269	5182	5395	5552
Chemicals and allied products	3875	4660	4930	5005	5114	5248	5630	5870	5876
Petroleum and coal products	4616	5181	5317	5337	5421	5710	6316	6455	6347
Rubber and plastic products	3538	3984	4014	4287	4221	4539	4721	4905	5481
Leather and Leather Products	3016	3196	3149	3317	3414	3254	3177	3346	4067
Average compensation	3396	3787	4013	4104	4146	4241	4387	4538	4843
Standard Deviation	743.7	851.4	727.8	746.3	929.9	1080.6	1243.2	1307.7	1112.5
Average Productivity	9355	8584	7962	8275	8389	9418	9795	10281	11750

Source: The Census of Manufactures, US Department of Commerce

Although the industries differ in their output composition, Table 5.1 shows clearly that on the average large firms pay more than the small firms. As an example, firms employing one to four employees paid on the average about $3396, whereas those employing 1000 to 2499 on the average paid about $4843. But note that the standard deviation of compensation also generally increases with the employment size of the establishment, suggesting that the higher the average pay, the higher its variability.

Suppose we want to run the regression

Average compensation$_i$ = $b_0 + b_1$ average productivity$_i$ + m_i where i refers to the ith employment-size class. If we were to run the preceding regression using the data given in Table 5.1, we would most likely encounter heteroscedasticity. Of course, we would have to find out whether the standard deviations of compensation presented in Table 5.1 are statistically significantly different.

Detection of Heteroscedasticity

The important practical question is: How does one know that heteroscedasticity is present in a specific situation? There are no hard and fast rules for detecting heteroscedasticity, only a few rules of thumb. But this is inevitable because σ_i^2 can be known only if we have the entire Y population corresponding to the chosen X's

such as the population shown in Table 5.1. But such data are an exception rather than the rule in most economic investigations. In this respect the econometrician differs from scientists in fields such as agriculture and biology where they have a good deal of control over their subjects. More often than not, in economic studies there is only one sample Y value corresponding to a particular value of X. And there is no way one can know σ_i^2 from just one Y observation.

Therefore, in most cases involving econometric investigations, heteroscedasticity may be a matter of 'speculation' or, as one author puts it, 'ad-hockery'. (The term is due to professor Zvi Griliches).

With the preceding caveat in mind, let us examine some of the informal and formal methods of detecting heteroscedasticity.

1. **Nature of the problem:** Very often the nature of the problem under consideration suggests whether heteroscedasticity is likely to be cncountered. For example, the pioneering work of Prais and Houthakker on family budget studies, wher they found that the residual variance around the regression of consumption on income increased with income, it is now generally assumed that in similar surveys one can expect unequal variances among the disturbances. As a matter of fact, in cross-sectional data involving heterogeneous units, heteroscedasticity may be the rule rather than the exception. Thus, in a cross-sectional analysis involving the investment expenditure in relation to sales, rate of interest, etc. Heteroscedasticity is generally expected is small, medium, and large-size firms are sampled together.
2. **Graphical method:** If there is no a priori or empirical information about the nature of heteroscedasticity, in practice one can do the regression analysis on the assumption that there is no heteroscedasticity and then do a post mortem examination of the estimated residual squared σ_i^2 to see if they exhibit any systematic pattern. Although σ_i^2 are not the same thing as μ_i^2, they can be used as proxies especially if the sample size is sufficiently large. An examination of the RSS may reveal patterns such as those shown in Figure 5.5.

NOTES

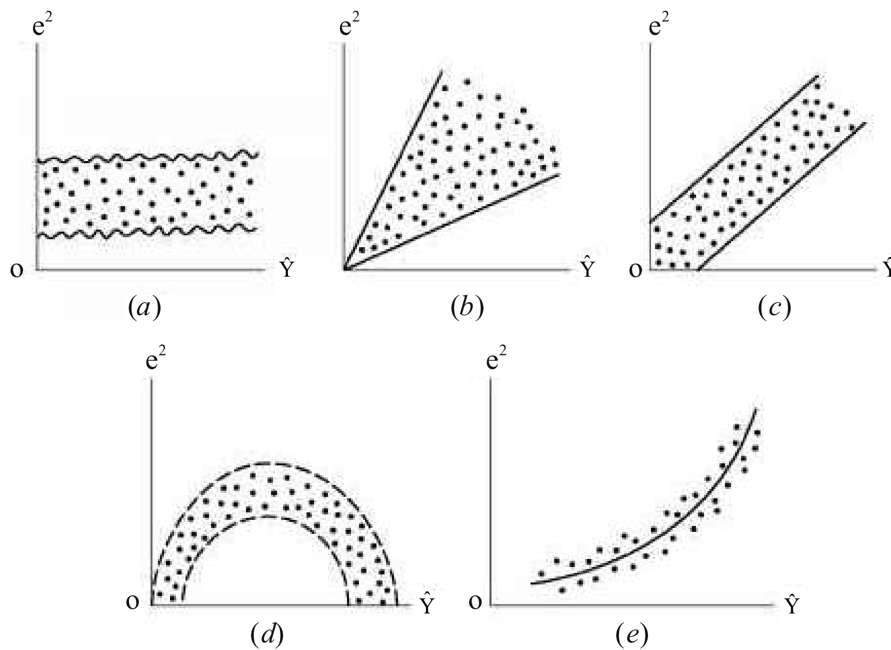


Fig. 5.5 *Heteroscedasticity*

In Figure 5.5, σ_i^2 are plotted $\hat{Y}_i$, the estimated Y_i from the regression line, the idea being to find out whether the estimated mean value of Y is systematically related to the squared residual. In Figure 5.5a we see that there is no systematic pattern between the two variables, suggesting that perhaps no heteroscedasticity is present in the data. Figure 5.5b to e, however, exhibits definite patterns. For instance, Figure 5.5c suggests a linear relationship whereas Figure 5.5d and e indicate a quadratic relationship between σ_i^2 and $\hat{Y}_i$. Using such knowledge, albit informal, one may transform the data in such a manner that the transformed data do not exhibit heteroscedasticity.

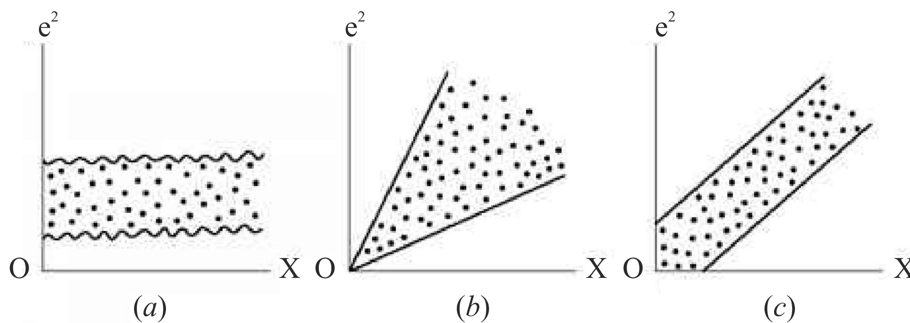


Fig. 5.6 *Regression Line Pattern*

Instead by plotting σ_i^2 against $\hat{Y}_i$, one may plot them against one of the explanatary variables especially if plotting σ_i^2 against $\hat{Y}_i$ results in the pattern shown in Figure 5.5a. Such a plot, which in shown in Figure 5.6, may reveal patterns similar to those given in Figure 5.5. (In the case of the two-variable model, plotting σ_i^2 against $\hat{Y}_i$ is equivalent to plotting it against X_i, and therefore Figure 5.6 is similar to Figure 5.5. But this is not the situation when we consider a model involving two or more X variables; in this instance, σ_i^2 may be plotted against any X variable included in the model.)

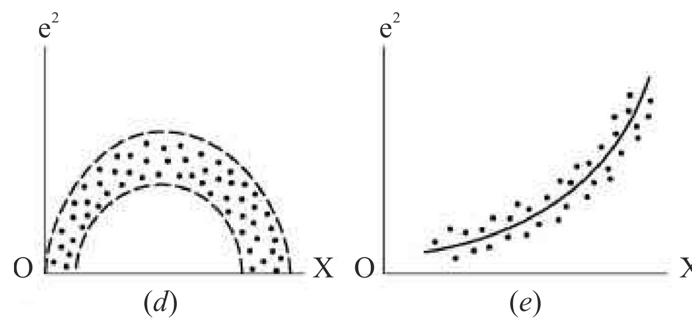


Fig. 5.7 *Variance of the Distribution*

A pattern such as that shown in Figure 5.6c, for instance, suggests that the variance of the distribution term is linearly related to the X variable. Thus, if in the regression of savings on income, one finds a pattern such as that shown in Figure 5.6c, it suggests that the heterascedastic variance may be proportional to the value of the income variable. This knowledge may help us in transforming our data in such a manner that in the regression on the transformed data the variance of the disturbance is homoscedastic. We shall return to this topic in the next section.

3. **Park test:** Park formalizes the graphical method by suggesting that σ_i^2 is some function of the explanatory variables X_i. The functional form he suggested was

$$\sigma_i^2 = \sigma^2 X_i^{\beta} e^{Vi}$$

or $$\text{In } \sigma_i^2 = \text{In } \sigma^2 + \beta \text{ In } X_i + V_i \qquad ...(5.3)$$

where V_i is the stochastic disturbance term.

Since σ_i^2 is generally not known, Park suggests using σ_i^2 as a proxy and running the following regression:

$$\text{In}\,\sigma_i^2 = \text{In } \sigma^2 + \beta \text{ In } X_i + V_i$$
$$= \alpha + \beta \text{ In } X_i + V_i \qquad ...(5.4)$$

If β turns out to be statistically significant, it would suggest that heteroscedasticity is present in the data. If it turns out to be insignificant, we may accept the assumption of homoscedasticity. The Park test is thus a

two-stage procedure. In the first stage we run the OLS regression and then in the second stage we run the regression (Equation 5.4).

Although empirically appealing, the Park test has some problems. Goldfeld and Quandt have argued that the error term v_i entering into (5.4) may not satisfy the OLS assumptions and may itself be heteroscedastic. Nonetheless, as a strictly exploratory method, one may use the Park test.

To illustrate the Park approach, we use the data given in Table 5.1 to run the following regression:

$$Y_i = \beta_0 + \beta_0 X_i + \mu_i$$

where Y = average compensation in thousands of dollars, X = average productivity in thousands of dollars, and i = ith employment size of the establishment. The results of the regression were as follows:

$$\hat{Y}_i = 1999.0466 + 0.2323X_i (0.1000) \qquad ...(5.5)$$

$$t = (2.323) \; R^2 = 0.4356$$

The results reveal that the estimated slope coefficient is significant at the 5 percent level on the basis of a one-tail *t* test. The equation shows that as labor productivity increases by, say, a dollar, labor compensation on the average increases by about 23 cents.

The residuals obtained from regression (5.5) were regressed on X_i as suggested in Equation 5.4, giving the following results:

$$\text{In}\,\sigma_i^2 = 35.9010 - 2.8099 \text{ In } X_i$$

$$(4.216) \qquad ...(5.6)$$

$$t = (-0.667) \quad R^2 = 0.0595$$

Obviously, there is no statistically significant relationship between the two variables. Following the Park test, one may conclude that there is no heteroscedasticity in the error variance.

4. **Glejser test:** The Glejser test is similar in spirit to the Park test. After obtaining the residuals e_i from the OLS regression, Glejser suggests regressing the absolute values of e_i, $|e_i|$ on the X variable that is thought to be closely associated with σ_i^2. In his experiments, Glejser used the following functional forms:

$$|e_i| = \beta_1 X_i + v_i$$

$$|e_i| = \beta_1 \sqrt{X_i} + v_i$$

$$|e_i| = \beta_1 \frac{1}{X_i} + v_i$$

$$|e_i| = \beta_1 \frac{1}{\sqrt{X_i}} + v_i$$

$$|e_i| = \beta_0 + \beta_1 X_i + v_i$$

NOTES

$$|e_i| = \beta_1 \frac{1}{\sqrt{\beta_0 + \beta_i X_i}} + v_i$$

$$|e_i| = \beta_1 \frac{1}{\sqrt{\beta_0 + \beta_i X_i^2}} + v_i$$

where v_i is the error term.

Again as an empirical or practical matter, one may use the Glejser approach. But Goldfeld and Quandt point out that the error term v_i has some problems in that its expected value is nonzero, it is serially correlated, and ironically it is heteroscedastic. An additional difficulty with the Glejser method is that models such as

$$|e_i| = \sqrt{\beta_0 + \beta_i X_i} + v_i \text{ and } |e_i| = \sqrt{\beta_0 + \beta_i X_i^2}$$

are nonlinear in the parameters and therefore cannot be estimated with the usual OLS procedure.

Glejser has found that for large samples the first four of the preceding models give generally satisfactory results in detecting heteroscedasticity. As a practical matter, therefore, the Glejser technique may be used for large samples and may be used in the small samples strictly as a qualitative device to learn something about heteroscedasticity.

6. **Spearman's rank correlation test:** The Spearman's rank correlation coefficient is defined as

$$r_s = 1 - 6\left[\frac{\Sigma d_i^2}{N(N^2 - 1]}\right] \qquad ...(5.7)$$

where d_i = difference in the ranks assigned to two different characteristics of the ith individual or phenomenon and N = number of individuals or phenomena ranked. The preceding rank correlation coefficient can be used to detect heteroscedasticity as follows: Assume $Y_i = \beta_0 + \beta_1 X_i + u_i$.

Step 1: Fit the regression to the data on Y and X and obtain the residuals e_i.

Step 2: Ignoring the sign of e_i, that is, taking their absolute value $|e_i|$, rank both $|e_i|$ and X_i according to an ascending or descending order and compute the Spearman's rank correlation coefficient given previously.

Step 3: Assuming that the population rank correlation coefficient P_s to be zero and $n > 8$, the significance of the sample r_s can be tested by the t test as follows:

$$t = \frac{r_s\sqrt{N-2}}{\sqrt{1-r_s^2}} \qquad ...(5.8)$$

with df = N – 2.

If the computed t value exceeds the critical t value, we may accept the hypothesis of heteroscedasticity; otherwise we may reject it. If the regression model

involves more than one X variable, r_s can be computed between $|e_i|$ and each of the X variables separately and can be tested for statistical significance by the t test given above.

To test this hypothesis, we apply the rank correlation technique. Table 5.2 gives the necessary data required in the analysis. Applying formula (5.7), we obtain

$$r_s = 1 - 6\ \frac{126.5}{10(100-1)}$$

$$= 0.2333 \qquad ...(5.9)$$

Table 5.2 *Illustration of Rank Correlation Method*

X, standard deviation of annual returns	|ei|, Absolute Value of residual	Rank of X	Rank of $\|e_i\|$	d	d^2
12.4	1.01	5	9	–4	16
14.4	1.26	7	10	–3	9
14.6	0.18	8	4	4	16
16.0	0.20	9	5	4	16
11.3	0.22	3.5	6	–2.5	6.25
10.0	0.60	1	7	–6	36
16.2	0.90	10	8	2	4
10.4	0.11	2	3	–1	1
13.1	0.07	6	2	4	16
11.3	0.03	3.5	1	2.5	6.25
				0	126.5

Notes: (a) The values of X and |e| are ranked in ascending order. (b) Since two of X values are identical, their rank is tied.

Applying the t test given in (5.8), we obtain

$$t = \frac{(0.2333)(\sqrt{8})}{\sqrt{1-0.0544}}$$

$$= 0.6786 \qquad ...(5.10)$$

For 8 df this t value is not significant even at the 10 per cent level of significance. Thus, there is no evidence of a systematic relationship between the explanatory variable and the absolute values of the residuals, which might suggest that there is no heteroscedasticity.

In addition to the tests discussed previously, one may use Bartlett's homogenity-of-variance test. But this test requires that we have data of the type given in Table 5.1, which provides us with reveral estimates of the variance of the phenomenon under consideration.

Remedial Measures

Heteroscedasticity does not destroy the unbiasedness and consistency properties of the OLS estimators, but they are no longer efficient, not even asymptotically (i.e.,

large sample size). This lack of efficiency makes the usual hypothesis-testing procedure of dubious value. Therefore, remedial measures may are clearly called for. There are two approaches to remediation: when σ_i^2 i is known and when σ_i^2 is not known.

When σ_1^2 is known: The Method of Weighted Least Squares

When σ_i^2 is known or can be estimated, the most straightforward method of dealing heteroscedasticity is by means of the weighted least squares. To illustrate this method, consider the two variable model.

$$\text{PRF: } Y_i = \beta_0 + \beta_1 X_i + \mu_i$$

$$\text{SRF: } Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

The usual or unweighted least squares method consists in minimizing RSS: $\Sigma e_i^2 = \Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$ with respect to the unknowns. In minimizing this RSS, the unweighted least-squares method gives implicity the same weight to each σ_i^2. Thus, in the hypothetical scattergram of Figure 5.7, points A, B and C all have the same weight in computing Σe_i^2. Obviously, in this case, the σ_i^2 associated with point C will dominate the RSS.

The method of weighted least squares does take into account the extreme points, such as C in Figure 5.7, by minmizing, not the usual or unweighted RSS, but the following RSS:

$$\text{Min: } \Sigma e_i^2 = \Sigma w_i (Y_i - \beta^*{}_0 - \beta^*{}_1 X_i)^2 \qquad ...(5.11)$$

where, w_i, the weights, are some constant (nonstochastic) numbers and where β_0^* and β_1^* are weighted least-squares estimators. The w_i are chosen in such a manner that the extreme observations for example, (in Figure 5.7) receive smaller weights. If σ_i^2 is known, one can let

$$wi = \frac{1}{\sigma_i^2} \qquad ...(5.12)$$

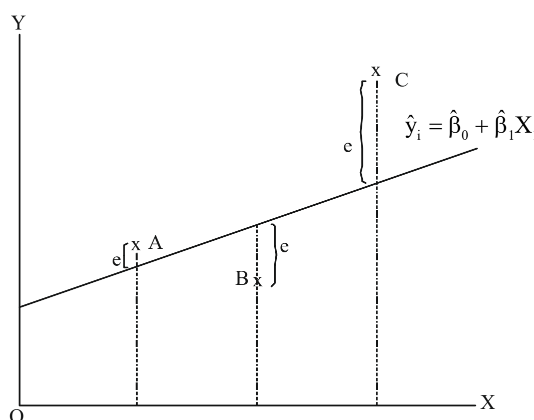


Fig. 5.8 *Method of Kleighted Least Squares*

that is, weight each observation inversely proportional to σ_i^2. This scheme of weighting will 'discount' heavily observations which come from populations with larger variances, such as point C in Figure 5.7

The mechanics of minimizing (5.11) follows the usual calculus techniques. The results of the minimization procedures are as follows:

$$\beta_0^* = \frac{\Sigma w_i Y_i}{\Sigma w_i} - \beta_1^* \frac{\Sigma w_i X_i}{\Sigma w_i}$$

$$= \bar{Y}^* - \beta_1^* \bar{X}^* \qquad ...(5.13)$$

where, $\bar{Y}^*$ and $\bar{X}^*$ are weighted sample means with w_i serving as the weights and $X_i^* = X_i \bar{X}^*$ represent deviations from the weighted sample means. If $w_1 = w_2 = w_3 = ... = w_N$, that is, each observation has the same weight, the weighted least-squares estimators given previously coincide with the usual or unweighted least-squares estimators.

When σ_1^2 is Not known

When σ_1^2 is not know, the method of weighted least squares discussed previously cannot be used readily. In practice, therefore, one may resort to some and hoc, albeit reasonably plausible, assumptions about σ_i^2 and transform the original regression model in such a way that the transformed model will satisfy the assumption of homoscedasticity. Without some such transformation, the problem of heteroscedasticity becomes practically insoluble. We now illustrate some of these transformations with the help of the two-variable model

$$Y_i = \beta_0 + \beta_2 X_i + \mu_i$$

Several possible assumptions about the pattern of heteroscedasticity are now considered.

Assumption 1

$$\in (\mu_i^2) = \sigma^2 X_i^2 \qquad (5.14)$$

If as a matter of "speculation," graphical methods, or Park and Glejser approaches it is believed that the variance of U_i is proportional to the square of the explanatory variable X, one may transform the original model as follows. Divide the original model through by X_i:

$$\frac{Y_i}{X_i} = \frac{\beta_0}{X_i} + \beta_1 + \frac{\mu_i}{X_i}$$

$$= \beta_0 \frac{1}{X_i} + \beta_1 + V_i \qquad (5.15)$$

Where V_i is the transformed disturbance term and is equal to μ_i / X_i. Now it is easy to verify that

NOTES

$$\in (V_i^2) = \in \left(\frac{\mu_2}{X_i} \right)^2 = \frac{1}{X_i^2} \in (\mu_i^2)$$

$$= \sigma^2$$

Hence the variance of V_i is homoscedastic, and one may proceed to apply OLS to the transformed equation (5.15), regressing Y_i/X_i on $1/X_i$.

Notice that in the transformed regression, the intercept term β_1 is the slope coefficient in the original equation and the slope coefficient β_1 is the intercept term in the original model. Therefore, to get back to the original model we shall have to multiply the estimated ( ) by X_i.

Assumption 2

$$\in (\mu_i^2) = \sigma^2 X_i \qquad (5.16)$$

If it is believed that the variance of μ_i instead of being proportional to the squared X_i is proportional to X_i itself, then the original model can be transformed as follows:

$$\frac{Y_i}{\sqrt{X_i}} = \frac{\beta_0}{\sqrt{X_i}} + \beta_1 \sqrt{X_i} + \frac{\mu_i}{\sqrt{X_i}}$$

$$= \beta_0 \frac{1}{\sqrt{X_i}} + \beta_1 \sqrt{X_i} + V_i \qquad (5.17)$$

where $V_i = U_i / \sqrt{X_i}$ and where $X_i > o$.

Given Assumption 2, it can be readily verified that $\in (V_i^2) = \sigma^2$, a homoscedastic situation. Therefore, one may proced to apply OLS to (5.18), regressing $Y_i / \sqrt{X_i}$ on $1/\sqrt{X_i}$ and $\sqrt{X_i}$.

Note that an important feature of the transformed model is that it has no intercept term. Therefore, one will have to use the "regression through the origin" model to estimate β0 and β1. Havin run (5.17), one can get back to the original model simply by multiplying (5.17) by $\sqrt{X_i}$

Assumption 3

$$\in (U_i^2) = \sigma^2 [\in (Y_i)]^2 \qquad (5.18)$$

NOTES

Equation (5.18) postulates that the variance of Ui is proportional to the square of the expected value of Y (see Fig. 5.5e). Now

$$\in (Y_i) = \beta_0 + \beta_1 X_i$$

Therefore, if we transform the original equation as follows:

$$\frac{Y_i}{\in (Y_i)} = \frac{\beta_0}{\in (Y_i)} + \beta_1 \frac{X_i}{\in (Y_i)} + \frac{U_i}{\in (Y_i)}$$

$$= \beta_0 \left(\frac{1}{\in (Y_i)} \right) + \beta_1 \frac{X_i}{\in (Y_i)} + V_i \qquad (5.19)$$

where $V_i = \mu_i / \in (Y_i)$, it can be seen that $V_i = \mu_i / \in (V_i^2) = \sigma^2$, that is, the disturbances V_i are homosedastic. Hence, it is regression (5.19) that will satisfy the homoscedasticity assumption of the classical linear regression model.

The transformation (5.19) is, however, inoperational because $\in(Y_i)$ depends on β_0 and β_1, which are unknown. Of course, we know $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i$, which is an estimate of $\in(Y_i)$. Therefore, we may proceed in two steps: first we run the usual OLS regression disregarding the heteroscedasticity problem and obtain $\hat{Y}_i$.

Then, using the estimated $\hat{Y}_i$, we transform our model as follows:

$$\frac{Y_i}{\hat{Y}_i} = \beta_0 \left(\frac{1}{\hat{Y}_i} \right) + \beta_1 \left(\frac{X_i}{\hat{Y}_i} \right) + V_i \qquad (5.20)$$

where $V_i = (\mu_i + \hat{Y}_i)$. In step 2, we run the regression (5.20). Although $\hat{Y}_i$ are not exactly $\in(Y_i)$, they are consistent estimators; that is, as the sample size increases indefinitely, they converge to true $\in(Y_i)$. Hence, the transformation ( ) will do in practice if the sample size is reasonably large.

Assumption 4 Log transformation: If, instead of running the regression $Y_i = \beta_0 + \beta_1 X_i + \mu_i$, we run

$$\text{In } Y_i = \beta_0 + \beta_1 In\, X_i + \mu_i \qquad (5.21)$$

Very often it reduces heteroscedasticity. This is because log transformation compresses the scales in which the variables are measured, thereby reducing a tenfold difference between two values to a twofold difference. Thus, the number 80 is 10 times, the number 8, but in 80 (=4.3820) is only twice as large as In 8 (=2.0794).

An additional advantages of the log transformation is that the slope coefficient β_1 measures the elasticity of Y with respect to X, that is, the percentage change in Y for a percentage change in X. For example, if Y is consumption and X is income

NOTES

β_1 in (5.21) will measure income elasticity, whereas in the original model β_1 measures only the rate of change of mean consumption for a unit change in income. It is one reason why the log models are quite popular in empirical econometrics.

To conclude our discussion of the remedial measures, it should be reemphasized that all the transformations discussed previously are ad hoc; we are essentially speculating about the nature of σ_i^2. Which of the transformations discussed previously will work will depend on the nature of the problem and the severity of heteroscedasticity. There are some additional problems with the transformations we have considered. For example, when we go beyond the two-variable model we may not know a priori which of the X variables should be chosen for transforming the data. Then there is the problem of spurious correlation. This term, due to Karl Pearson, refers to the situation where correlation is found to be present between the ratios of variables even though the original variables are uncorrelated or random. Thus, in the model $Y_i = \beta_0 + \beta_1 X_i + \mu_i$, Y and X may not be correlated; but in the transformed model

$Y_i / X_i = \beta_0 (1/X_i) + \beta_1 + V_i, Y_i / X_i$ are often found to be correlated. Therefore, the reader should be aware of some of the problems associated with the commonly used transformations in econometric studies.

5.5 SERIAL CORRELATIONS

In econometrics, serial correlation is used to describe the relationship between the observations of the same variable over the specific time periods. If a variable's serial correlation is measured as zero, there is no correlation, and each of the observations is independent of one another. Hence, serial correlation happens in a time series when a variable and a lagged version of itself are observed to be correlated with one another, over the periods of time. Although, the term "Serial Correlation" or simply "Autocorrelation" may be used to describe a relation of any variable, but in traditional econometrics, it is meant to refer to a particular variable or time series: that of the errors of the regression model.

In econometrics and statistics, the autocorrelation of a real or complex random process is the Pearson correlation between values of the process at different times, as a function of the two times or of the time lag. Let $\{X_t\}$ be a random process, and t be any point in time (may be an integer for a discrete-time process or a real number for a continuous-time process). Then $\{X_t\}$ is the value (or realization) produced by a given run of the process at time *t*. Suppose that the process has mean μ_t and variance σ^2_t at time *t*, or each *t*. Then the definition of the auto-correlation function between times t_1 and t_2 is:

$$\mathrm{R}_{XX}(t_1, t_2) = \mathrm{E}\left[X_{t_1} \overline{X}_{t_2}\right] \tag{5.22}$$

Where the expected value operator and the bar is represents complex conjugation. Note that the expectation may not be well defined.

Subtracting the mean before multiplication yields the auto-covariance function between times t_1 and t_2

$$K_{XX}(t_1, t_2) = E\left[(X_{t_1} - \mu_{t_1})\overline{(X_{t_2} - \mu_{t_2})}\right] = E\left[X_{t_1}\overline{X}_{t_2}\right] - \mu_{t_1}\overline{\mu}_{t_2} \quad (5.23)$$

Note that this expression is not well-defined for all-time series or processes, because the mean may not exist, or the variance may be zero (for a constant process) or infinite (for processes with distribution lacking well-behaved moments, such as certain types of power law).

Definition for Wide-Sense Stationary Stochastic Process

If $\{X_t\}$ is a wide-sense stationary process than, the mean μ and the variance σ^2 are time-independent, and further the auto covariance function depends only on the lag between t_1 and t_2: the auto covariance depends only on the time-distance between the pair of values but not on their position in time. This further implies that the auto covariance and auto-correlation can be expressed as a function of the time-lag, and that this would be an even function of the lag $\tau = t_2 - t_1$. This gives the more familiar forms for the auto-correlation function

$$R_{XX}(\tau) = E\left[X_t\overline{X}_{t+\tau}\right] \quad (5.24)$$

And the auto-covariance function:

$$K_{XX}(\tau) = E\left[(X_t - \mu)\overline{(X_{t+\tau} - \mu)}\right] = E\left[X_t\overline{X}_{t+\tau}\right] - \mu\overline{\mu} \quad (5.25)$$

In regression analysis using time series data, autocorrelation in a variable of interest is typically modelled either with an Auto Regressive model (AR), a Moving Average model (MA), their combination as an autoregressive-moving-average model (ARMA), or an extension of the latter called an Auto Regressive Integrated Moving Average model (ARIMA). With multiple interrelated data series, Vector Auto Regression (VAR) or its extensions are used.

In Ordinary Least Squares (OLS), the adequacy of a model specification can be checked in part by establishing whether there is autocorrelation of the regression residuals. Problematic autocorrelation of the errors, which themselves are unobserved, can generally be detected because it produces autocorrelation in the observable residuals. (Errors are also known as "Error Terms" in econometrics.) Autocorrelation of the errors violates the ordinary least squares assumption that the error terms are uncorrelated, meaning that the Gauss Markov theorem does not apply, and that OLS estimators are no longer the Best Linear Unbiased Estimators (BLUE). While it does not bias the OLS coefficient estimates, the standard errors tend to be underestimated (and the t-scores overestimated) when the autocorrelations of the errors at low lags are positive.

NOTES

Responses to nonzero autocorrelation include generalized least squares and the Newey–West HAC estimator (Heteroskedasticity and Autocorrelation Consistent). Autocorrelation is used to analyse dynamic light scattering data, which notably enables determination of the particle size distributions of nanometre-sized particles or micelles suspended in a fluid. A laser shining into the mixture produces a speckle pattern that results from the motion of the particles. Autocorrelation of the signal can be analysed in terms of the diffusion of the particles. From this, knowing the viscosity of the fluid, the sizes of the particles can be calculated.

Check Your Progress

1. Explain the violations of classical assumptions consequences.
2. Define the term multicollinearity.
3. Illustrate the effects of multicollinearity.
4. State the detecting multicollinearity.
5. Interpret the remedies for multicollinearity.
6. Elaborate on the heteroscedasticity.
7. Explain the detection of heteroscedasticity.
8. Define the Park test.
9. What do you understand by the Glejser test?
10. Interpret the Spearman's rank correlation test.

5.6 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. There are several different frameworks in which the linear regression model can be cast in order to make the OLS technique applicable. Each of these settings produces the same formulas and same results. The only difference is the interpretation and the assumptions which have to be imposed in order for the method to give meaningful results.
2. Multicollinearity refers to the phenomenon where, in a multiple regression model, two or a greater number of predictor variables are variables are highly correlated. This implies that one can be linearly predicted from the others with accuracy of a high degree.
3. An effect of high degree of multicollinearity is that, despite matrix X^TX being invertible, the computer algorithm might not be successful in getting an approximate inverse, and in case it is obtained by the algorithm, it might lack numerical inaccuracy. Yet, even if the X^TX matrix is accurate, there is the possibility of several effects.

4. It has been suggested that there be a formal detection-tolerance or the Variance Inflation Factor (VIF) for multicollinearity:

$$\text{tolerance} = 1 - R_j^2, \quad \text{VIF} = \frac{1}{\text{tolerance}}$$

5. Stay away from the dummy variable trap; including a dummy variable for every category and including a constant term in the regression will ensure perfect multicollinearity.
6. The word 'Heteroscedasticity' comes from the Greek, and quite literally means data with a different (hetero) dispersion (skedasis). In simple terms heteroscedasticity is any set of data that isn't homoscedastic. More technically, it refers to data with unequal variability (scalter) across a set of second, prefictor variables.
7. The important practical question is: How does one know that heteroscedasticity is present in a specific situation? There are no hard and fast rules for detecting heteroscedasticity, only a few rules of thumb. But this is inevitable because σ_i^2 can be known only if we have the entire Y population corresponding to the chosen X's.
8. Park formalizes the graphical method by suggesting that σ_i^2 is some function of the explanatory variables X_i. The functional form he suggested was

   $\sigma_i^2 = \sigma^2 X_i^{\beta} e^{vi}$

   Or In σ_i^2 = In $\sigma^2 + \beta$ In $X_i + V_i$, where V_i is the stochastic disturbance term.
9. The Glejser test is similar in spirit to the Park test. After obtaining the residuals e_i from the OLS regression, Glejser suggests regressing the absolute values of e_i, $|e_i|$ on the X variable that is thought to be closely associated with σ_i^2.
10. The Spearman's rank correlation coefficient is defined as

$$r_s = 1 - 6\left[\frac{\Sigma d_i^2}{N(N^2 - 1]}\right] \quad .$$

5.7 SUMMARY

- There are several different frameworks in which the linear regression model can be cast in order to make the OLS technique applicable. Each of these settings produces the same formulas and same results.
- The only difference is the interpretation and the assumptions which have to be imposed in order for the method to give meaningful results. The choice of the applicable framework depends mostly on the nature of data in hand, and on the inference task which has to be performed.
- The classical model focuses on the "Finite Sample" estimation and inference, meaning that the number of observations n is fixed. This contrasts with the other approaches, which study the asymptotic behaviour of OLS, and in which the number of observations is allowed to grow to infinity.

NOTES

- The exogeneity assumption is critical for the OLS theory. If it holds then the regressor variables are called exogenous. If it doesn't, then those regressors that are correlated with the error term are called endogenous, and then the OLS estimates become invalid.
- One consequence of a high degree of multicollinearity is that, even if the matrix $X^{\top} X$ is invertible, a computer algorithm may be unsuccessful in obtaining an approximate inverse, and if it does obtain one it may be numerically inaccurate. But even in the presence of an accurate matrix, the following consequences arise.
- Multicollinearity refers to the phenomenon where, in a multiple regression model, two or a greater number of predictor variables are variables are highly correlated. This implies that one can be linearly predicted from the others with accuracy of a high degree.
- An effect of high degree of multicollinearity is that, despite matrix X^TX being invertible, the computer algorithm might not be successful in getting an approximate inverse, and in case it is obtained by the algorithm, it might lack numerical inaccuracy. Yet, even if the X^TX matrix is accurate, there is the possibility of several effects.
- It has been suggested that there be a formal detection-tolerance or the Variance Inflation Factor (VIF) for multicollinearity:

$$\text{tolerance} = 1 - R_j^2, \quad \text{VIF} = \frac{1}{\text{tolerance}}$$

- The word 'Heteroscedasticity' comes from the Greek, and quite literally means data with a different (hetero) dispersion (skedasis). In simple terms heteroscedasticity is any set of data that isn't homoscedastic. More technically, it refers to data with unequal variability (scalter) across a set of second, prefictor variables.
- Park formalizes the graphical method by suggesting that σ_i^2 is some function of the explanatory variables X_i. The functional form he suggested was

  $\sigma_i^2 = \sigma^2 X_i^{\beta} e^{vi}$

  Or $\text{In } \sigma_i^2 = \text{In } \sigma^2 + \beta \text{ In } X_i + V_i$, where V_i is the stochastic disturbance term.
- The Glejser test is similar in spirit to the Park test. After obtaining the residuals e_i from the OLS regression, Glejser suggests regressing the absolute values of e_i, $|e_i|$ on the X variable that is thought to be closely associated with σ_i^2.
- The Spearman's rank correlation coefficient is defined as

$$r_s = 1 - 6\left[\frac{\Sigma d_i^2}{N(N^2 - 1]}\right].$$

- In econometrics, serial correlation is used to describe the relationship between the observations of the same variable over the specific time periods. If a variable's serial correlation is measured as zero, there is no correlation, and each of the observations is independent of one another.

- In econometrics and statistics, the autocorrelation of a real or complex random process is the Pearson correlation between values of the process at different times, as a function of the two times or of the time lag. In Ordinary Least Squares (OLS), the adequacy of a model specification can be checked in part by establishing whether there is autocorrelation of the regression residuals. Problematic autocorrelation of the errors, which themselves are unobserved, can generally be detected because it produces autocorrelation in the observable residuals.

5.8 KEY WORDS

- **Classical model:** The classical model focuses on the "Finite Sample" estimation and inference, meaning that the number of observations n is fixed.
- **Exogeneity assumption:** The exogeneity assumption is critical for the OLS theory. If it holds then the regressor variables are called exogenous.
- **Collinear variables:** The collinear variables contain the same information about the dependent variable. If nominally "Different" measures actually quantify the same phenomenon then they are redundant.
- **Multicollinearity:** Multicollinearity refers to the phenomenon where, in a multiple regression model, two or a greater number of predictor variables are variables are highly correlated.
- **Heteroscedasticity:** The word 'Heteroscedasticity' comes from the Greek, and quite literally means data with a different (hetero) dispersion (skedasis). In simple terms heteroscedasticity is any set of data that isn't homoscedastic. More technically, it refers to data with unequal variability (scalter) across a set of second, prefictor variables.
- **Glejser test:** The Glejser test is similar in spirit to the Park test. After obtaining the residuals e_i from the OLS regression, Glejser suggests regressing the absolute values of e_i, $|e_i|$ on the X variable that is thought to be closely associated with σ_i^2.
- **Serial correlation:** serial correlation is used to describe the relationship between the observations of the same variable over the specific time periods. If a variable's serial correlation is measured as zero, there is no correlation, and each of the observations is independent of one another.

5.9 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Define the violations of classical assumptions consequences.
2. Explain the term multicollinearity.

3. Interpret the effects of multicollinearity.
4. State the detecting multicollinearity.
5. Illustrate the remedies for multicollinearity.
6. What do you understand by the heteroscedasticity?
7. Define the detection of heteroscedasticity.
8. Explain the Park test.
9. Elaborate on the Glejser test.
10. State the Spearman's rank correlation test.

Long-Answer Questions

1. Briefly define the violations of classical assumptions consequences with the help of examples.
2. Discuss the multicollinearity. Explain the effects of multicollinearity.
3. Explain the detecting and remedies for multicollinearity.
4. Describe the heteroscedasticity. What do you mean by the detection of heteroscedasticity? Give appropriate examples.
5. Analyse the Park and Glejser test.
6. Elaborate on the Spearman's rank correlation test.

5.10 FURTHER READINGS

Johnston, J. and John DiNARDO. 1997. *Econometric Methods*, Fourth Edition. New Delhi: Tata McGraw-Hill.

Koutsoyiannis, A.1977. *Theory of Econometrics,* Second Edition. London: The Macmillan Press Ltd.

Özdemir, Durmuº. 2016. *Applied Statistics for Economics and Business,* Second Edition. Izmir (Turkey): Springer.

Maddala, G. S. 1992. *Introduction to Econometrics,* Second Edition. New York: Macmillan Publishing Company.

Pindyck, R. S and D. L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts,* Fourth Edition. New York: McGraw Hill.

Goldberger, A. S. 1998. *Introductory Econometrics***.** Cambridge: Harvard University Press.

Levine, David M., Timothy C. Krehbiei, Mark L. Berenson and P. K. Viswanathan. 2009. *Business Statistics*, Fifth Edition. New Delhi: Pearson Education.

Webster, Allen L. 1998. *Applied Statistics for Business and Economics,* Third Edition. New Delhi: Tata McGraw-Hill.

UNIT 6 SPECIFICATION ANALYSIS

NOTES

Structure

6.0 INTRODUCTION

Specification analysis plays an important role in econometrics. Model specification is part of the process of building a statistical model: specification consists of selecting an appropriate functional form for the model and choosing which variables to include. For example, given personal income y together with years of schooling s and on-the-job experience x, we might specify a functional relationship $y = f(s, x)$ as follows: $\ln y = \ln y_0 + \rho s + \beta_1 x + \beta_2 x^2 + \varepsilon$. Where ε is the unexplained error term that is supposed to comprise independent and identically distributed Gaussian variables. The statistician Sir David Cox has said, "How the translation from subject-matter problem to statistical model is done is often the most critical part of an analysis"?

Specification error occurs when the functional form or the choice of independent variables poorly represent relevant aspects of the true data-generating process. In particular, bias (the expected value of the difference of an estimated parameter and the true underlying value) occurs if an independent variable is correlated with the errors inherent in the underlying process. There are several different possible causes of specification error; some are as: An inappropriate functional form could be employed, a variable omitted from the model may have a relationship with both the dependent variable and one or more of the independent variables (causing omitted-variable bias), An irrelevant variable may be included in the model (although this does not create bias, it involves over fitting and so can lead to poor predictive performance), and the dependent variable may be part of a system of simultaneous equations (giving simultaneity bias).

NOTES

The Durbin–Wu–Hausman test (also called Hausman specification test) is a statistical hypothesis test in econometrics named after James Durbin, De-Min Wu, and Jerry A. Hausman. The test evaluates the consistency of an estimator when compared to an alternative, less efficient estimator which is already known to be consistent. It helps one evaluate if a statistical model corresponds to the data. The Ramsey Regression Equation Specification Error Test (RESET) test is a general specification test for the linear regression model. More specifically, it tests whether non-linear combinations of the fitted values help explain the response variable. The intuition behind the test is that if non-linear combinations of the explanatory variables have any power in explaining the response variable, the model is misspecified in the sense that the data generating process might be better approximated by a polynomial or another non-linear functional form.

In this unit, you will study about the specification analysis, omission of a relevant variable, inclusion of irrelevant variable, and test of specification errors.

6.1 OBJECTIVES

After going through this unit, you will be able to:

- Define the specification analysis
- Understand the omission of a relevant variable
- Explain the inclusion of irrelevant variable
- Analyse the test of specification errors

6.2 BASIC CONCEPT OF SPECIFICATION ANALYSIS

In econometrics, specification tests plays an important role to verify the validity of one specification at a time. It is said that most of these tests are not, in general, strong in the presence of other misspecifications. These tests will 'Confirm' the validity (or invalidity) of a general model requiring the estimates of the restricted model only.

Model specification is part of the process of building a statistical model: specification consists of selecting an appropriate functional form for the model and choosing which variables to include. For example, given personal income y together with years of schooling s and on-the-job experience x, we might specify a functional relationship $y = f(s, x)$ as follows:

$$\ln y = \ln y_0 + \rho s + \beta_1 x + \beta_2 x^2 + \varepsilon \cdot$$

Where ε is the unexplained error term that is supposed to comprise independent and identically distributed Gaussian variables. The statistician Sir David

Cox has said, "How [the] translation from subject-matter problem to statistical model is done is often the most critical part of an analysis"?

Specification error occurs when the functional form or the choice of independent variables poorly represent relevant aspects of the true data-generating process. In particular, bias (the expected value of the difference of an estimated parameter and the true underlying value) occurs if an independent variable is correlated with the errors inherent in the underlying process. There are several different possible causes of specification error; some are as: An inappropriate functional form could be employed, a variable omitted from the model may have a relationship with both the dependent variable and one or more of the independent variables (causing omitted-variable bias), An irrelevant variable may be included in the model (although this does not create bias, it involves over fitting and so can lead to poor predictive performance), and the dependent variable may be part of a system of simultaneous equations (giving simultaneity bias).

The Durbin–Wu–Hausman test (also called Hausman specification test) is a statistical hypothesis test in econometrics named after James Durbin, De-Min Wu, and Jerry A. Hausman. The test evaluates the consistency of an estimator when compared to an alternative, less efficient estimator which is already known to be consistent. It helps one evaluate if a statistical model corresponds to the data.

The Ramsey Regression Equation Specification Error Test (RESET) test is a general specification test for the linear regression model. More specifically, it tests whether non-linear combinations of the fitted values help explain the response variable. The intuition behind the test is that if non-linear combinations of the explanatory variables have any power in explaining the response variable, the model is misspecified in the sense that the data generating process might be better approximated by a polynomial or another non-linear functional form.

One approach is to start with a model in general form that relies on a theoretical understanding of the data-generating process. Then the model can be fit to the data and checked for the various sources of misspecification, in a task called statistical model validation. Theoretical understanding can then guide the modification of the model in such a way as to retain theoretical validity while removing the sources of misspecification. But if it proves impossible to find a theoretically acceptable specification that fits the data, the theoretical model may have to be rejected and replaced with another one.

A quotation from Karl Popper is apposite here: "Whenever a theory appears to you as the only possible one, take this as a sign that you have neither understood the theory nor the problem which it was intended to solve". Another approach to model building is to specify several different models as candidates, and then compare those candidate models to each other. The purpose of the comparison is to determine which candidate model is most appropriate for statistical inference.

Common criteria for comparing models include the following: R2, Bayes factor, and the likelihood-ratio test together with its generalization relative likelihood.

For more on this topic, see statistical model selection. The specification of a linear regression model consists of a formulation of the regression relationships and of statements or assumptions concerning the explanatory variables and disturbances. The complete regression analysis depends on the explanatory variables present in the model.

6.3 OMISSION OF A RELEVANT VARIABLE

The analyst may delete some of the explanatory variables to keep the model simple. There can be numerous causes behind such decisions, e.g., it may be hard to quantify the variables like the taste, intelligence, etc. Sometimes, it may be difficult to take correct observations on the variables like income etc.

Let there be k candidate explanatory variables out of which suppose r variables are included and $(k \times r)$ variables are to be deleted from the model. So partition the X and β as

$$\underset{n\times k}{X} = \begin{pmatrix} \underset{n\times r}{X_1} & \underset{n\times(k-r)}{X_2} \end{pmatrix} \text{ and } \beta = \begin{pmatrix} \underset{r\times 1}{\beta_1} & \underset{(k-r)\times 1)}{\beta_2} \end{pmatrix}.$$

The model $y = X\beta + \varepsilon,\ E(\varepsilon) = 0,\ V(\varepsilon) = \sigma^2 I$ can be expressed as

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

Which is called a full model or true model.

After dropping the r explanatory variable in the model, the new model is

$$y = X_1\beta_1 + \delta$$

Which is called a misspecified model or false model.

Applying OLS to the false model, the OLSE of β_1 is

$$b_{1F} = (X_1'X_1)^{-1}X_1'y.$$

The estimation error is obtained as follows:

$$b_{1F} = (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon)$$
$$= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon$$
$$b_{1F} - \beta_1 = \theta + (X_1'X_1)^{-1}X_1'\varepsilon$$

Where $\theta = (X_1'X_1)^{-1}X_1'X_2\beta_2$.

Thus,

$$E(b_{1F} - \beta_1) = \theta + (X_1'X_1)^{-1}E(\varepsilon)$$
$$= \theta$$

NOTES

Which is a linear function of β_2, i.e., the coefficients of excluded variables. So b_{1F} is biased, in general. The bias vanishes if $X_1'X_2 = 0$, i.e., X_1 and X_2 are orthogonal or uncorrelated.

The mean squared error matrix of b_{1F} is

$$MSE(b_{1F}) = E(b_{1F} - \beta_1)(b_{1F} - \beta_1)$$

$$\begin{aligned}
&= E\left[\theta\theta' + \theta\varepsilon' X_1(X_1'X_1)^{-1} + (X_1'X_1)^{-1}X_1'\varepsilon\theta' + (X_1'X_1)^{-1}X_1'\varepsilon\varepsilon' X_1(X_1'X_1)^{-1}\right] \\
&= \theta\theta' + 0 + 0 + \sigma^2(X_1'X_1)^{-1}X_1'IX_1(X_1'X_1)^{-1} \\
&= \theta\theta' + \sigma^2(X_1'X_1)^{-1}.
\end{aligned}$$

So efficiency generally declines. Note that the second term is the conventional form of MSE.

The residual sum of squares is

$$\hat{\sigma}^2 = \frac{SS_{res}}{n-r} = \frac{e'e}{n-r}$$

Where $e = y - X_1 b_{1F} = H_1 y$,

$$\bar{H}_1 = I - X_1(X_1'X_1)^{-1}X_1'.$$

Thus,

$$\begin{aligned}
\bar{H}_1 y &= \bar{H}_1(X_1\beta_1 + X_2\beta_2 + \varepsilon) \\
&= 0 + \bar{H}_1(X_2\beta_2 + \varepsilon) \\
&= \bar{H}_1(X_2\beta_2 + \varepsilon).
\end{aligned}$$

$$\begin{aligned}
y'\bar{H}_1 y &= (X_1\beta_1 + X_2\beta_2 + \varepsilon)\bar{H}_1(X_2\beta_2 + \varepsilon) \\
&= (\beta_2'X_2\bar{H}_1'\bar{H}_1X_2\beta_2 + \beta_2'X_2'\bar{H}_1\varepsilon + \beta_2'X_2\bar{H}_1'X_2\beta_2 + \beta_1'X_1'\bar{H}_1\varepsilon + \varepsilon'\bar{H}_1'X_2\beta_2 + \varepsilon'\bar{H}_1\varepsilon).
\end{aligned}$$

$$\begin{aligned}
E(s^2) &= \frac{1}{n-r}\left[E(\beta_2'X_2'\bar{H}_1X_2\beta_2) + 0 + 0 + E(\varepsilon'\bar{H}\varepsilon)\right] \\
&= \frac{1}{n-r}\left[\beta_2'X_2'\bar{H}_1X_2\beta_2) + (n-r)\sigma^2\right] \\
&= \sigma^2 + \frac{1}{n-r}\beta_2'X_2'\bar{H}_1X_2\beta_2.
\end{aligned}$$

Thus s^2 is a biased estimator of σ^2 and s^2 provides an overestimate of σ^2. Note that even if $X_1'X_2 = 0$, then also s^2 gives an overestimate of σ^2. So the statistical inferences based on this will be faulty. The t –test and confidence region will be invalid in this case.

NOTES

6.4 INCLUSION OF IRRELEVANT VARIABLE

Some variables may contribute very little to the explanatory power of the model. This may tend to reduce the degrees of freedom ($n–k$), and consequently, the validity of inference drawn may be questionable. For example, the value of the coefficient of determination will increase, indicating that the model is getting better, which may not really be true.

Let the true model be

$$y = X\beta + \varepsilon,\ E(\varepsilon) = 0, V(\varepsilon) = \sigma^2 I$$

Which comprise k explanatory variable. Suppose now r additional explanatory variables are added to the model and resulting model becomes

$$y = X\beta + Z\gamma + \delta$$

Where Z is an $n\times r$ matrix of n observations on each of the r explanatory variables and γ is $r\times 1$ vector of regression coefficient associated with Z and δ is disturbance term. This model is termed as a false model.

Applying OLS to false model, we get

$$\begin{pmatrix} b_F \\ c_F \end{pmatrix} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}^{-1} \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix} \begin{pmatrix} b_F \\ c_F \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

$$\Rightarrow X'Xb_F + X'Zc_F = X'y \tag{6.1}$$

$$Z'Xb_F + Z'Zc_F = Z'y \tag{6.2}$$

Where b_F and C_F are the OLSEs of β and , respectively.

Premultiply Equation (6.2) by $X'Z(Z'Z)^{-1}$, we get

$$X'Z(Z'Z)^{-1}Z'Xb_F + X'Z(Z'Z)^{-1}Z'Zc_F = X'Z(Z'Z)^{-1}Z'y. \tag{6.3}$$

Subtracting Equation (6.1) from (6.3), we get

$$\left[X'X - X'Z(Z'Z)^{-1}Z'X\right]b_F = X'y - X'Z(Z'Z)^{-1}Z'y$$

$$X'\left[I - Z(Z'Z)^{-1}Z'\right]Xb_F = X'\left[I - Z(Z'Z)^{-1}Z'\right]y$$

$$\Rightarrow b_F = (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z y$$

Where $\bar{H}_Z = I - Z(Z'Z)^{-1}Z'$.

The estimation error of b_F is

$$\begin{aligned} b_F - \beta &= (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z y - \beta \\ &= (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z(X\beta + \varepsilon) - \beta \\ &= (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z\varepsilon. \end{aligned}$$

Thus,

$$E(b_F - \beta) = (X'\bar{H}_Z X)^{-1}X'\bar{H}_Z E(\varepsilon) = 0$$

So, b_F is unbiased even when some irrelevant variables are added to the model.

The covariance matrix is

$$\begin{aligned} V(b_F) &= E\left(b_F - \beta\right)\left(b_F - \beta\right)^1 \\ &= E\left[\left(X'\bar{H}_Z X\right)^{-1}X'\bar{H}_Z\varepsilon\varepsilon'\bar{H}_Z X(X'\bar{H}_Z X)^{-1}\right] \\ &= \sigma^2\left(X'\bar{H}_Z X\right)^{-1}X'\bar{H}_Z I\bar{H}_Z X\left(X'\bar{H}_Z X\right)^{-1} \\ &= \sigma^2\left(X'\bar{H}_Z X\right)^{-1}. \end{aligned}$$

6.5 TEST OF SPECIFICATION ERRORS

In a linear regression model, the process of model specification requires three key decisions:

- Selection of the independent variables
- Omitting of the variables
- Selecting the functional form

The question arises what will be the impact when the linear regression model is used without meeting all the assumptions, that is, when it is not appropriate? Moreover, what properties do OLS estimators have when there is a specification error?

NOTES

Let us look at the following specification errors and what impact they can have. The specification errors are as follows:

(a) Omission of relevant variables

(b) Inclusion of irrelevant variables

(c) Measurement errors in variables

(a) Omitting a Relevant Variable

Let us begin by looking at the following multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Due to reasons like unavailability of the data on X_2, the regression model is created without variable X_2 as follows:

$$Y = \gamma_0 + \gamma_1 X_1 + \varepsilon_1$$

This issue, in econometrics, is called 'omitting a relevant variable'. To take an example, if $\beta_2 \neq 0$, and this is a type of misspecification, then the following equation shows what will be the consequence of omitting a relevant variable.

$$\gamma_1 = \beta_1 + \beta_2 \frac{C(X_1, X_2)}{V(X_1)} \tag{6.4}$$

The above equation (6.4) is referred to as the 'rule of omitted variable'. This depicts that the slope of the reduced model is a linear combination of β_1 and β_2 (2 slopes of the full model). Further, with omitting X_2, the effect will be part of the error term in the reduced model as follows:

$$\varepsilon_1 = \varepsilon + \beta_2 X_2$$

This implies the following:

$$\begin{aligned} E[\varepsilon_1|X_1] &= E[\varepsilon + \beta_2 X_2|X_1] \\ &= E[\varepsilon|X_1] + E[\beta_2 X_2|X_1] \\ &= E[E[\varepsilon|X_1, X_2]|X_1] + \beta_2 E\,[X_2|X_1] \\ &= \beta_2 E\,[X_2|X_1] \end{aligned}$$

Therefore,

$$E[Y \mid X_1] = \gamma_0 + \gamma_1 X_1 + \beta_2 E[X_2 \mid X_1]$$

Properties of the OLS Estimators

This implies that the model is one which is misspecified for a multiple linear model:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

The following exists:

$$E[\varepsilon \mid X_1, \ldots, X_k] \neq 0$$

To consider the OLS estimators' properties when there is such a specification error, consider the simplest multiple linear regression, and a simple regression using only one of the variables.

The MLR model is considered to be:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

And its estimated model is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

The SR model is considered to be:

$$Y = \gamma_0 + \gamma_1 X_1 + \varepsilon,$$

And its estimated model is:

$$\hat{Y} = \hat{\gamma}_0 + \hat{\gamma}_1 X_1$$

Thus, it is evident that:

$$\hat{\gamma}_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1$$

where $\hat{\delta}_1$ is the OLS estimator of the slope of $L(X_2 \mid X_1) = \delta_0 + \delta_1 X_1$.

The following is also known:

$$E\left[\hat{\beta}_1\right] = \beta_1 \text{ and } \operatorname{plim} \hat{\beta}_1 = \beta_1$$

$$E\left[\hat{\gamma}_1\right] = \gamma_1 \text{ and } \operatorname{plim} \hat{\gamma}_1 = \gamma_1$$

Putting together all of the above information, the following can be derived:

$$E\left[\hat{\gamma}_1 \mid X_1, X_2\right] = E\left[\hat{\beta}_1 \mid X_1, X_2\right] + E\left[\hat{\beta}_2 \hat{\delta}_1 \mid X_1, X_2\right] = \beta_1 + \beta_2 \hat{\delta}_1$$

This implies that:

$$E\left[\hat{\gamma}_1\right] = \beta_1 + \beta_2 E\left[\hat{\delta}_1\right]$$

$$\operatorname{plim}\left(\hat{\gamma}_1\right) = \beta_1 + \beta_2 \delta_1$$

Generally speaking, $\hat{\gamma}_1$ will not be appropriate in case the inference needs to be made regarding β_1. Also, it can be easily shown that:

$$V\left(\hat{\gamma}_1\right) \leq V\left(\hat{\beta}_1\right)$$

When X_2 is a 'relevant variable', implying that $\beta_2 \neq 0, \hat{\gamma}_1 \Rightarrow$ is a biased and inconsistent estimator of $\beta_1 \Rightarrow$, irrespective of the size of the sample, the bias will remain.

$V\left(\hat{\gamma}_1\right) = \dfrac{\sigma^2}{\sum_i x_{1_i}}$ and $V\left(\hat{\beta}_1\right) = \dfrac{\sigma^2}{\sum_i x_{1_i}\left(1 - R_1^2\right)}$, that is $V\left(\hat{\gamma}_1\right) \leq V\left(\hat{\beta}_1\right)$, but $\hat{V}\left(\hat{\gamma}_1\right)$ is a biased estimator for the variance of $\hat{\beta}_1$.

NOTES

This depicts that the 'omission of relevant variables' in the analysis causes bias and inconsistency when estimating the effects of variables, by reducing the estimator's variance.

That is to say, it is the coefficient of X_1 in the regression which incorrectly omits X_2, since it:

- captures what the effect is upon Y when there is a change in X1 and also the effect of X_1 on X_2 (which also affects Y)
- does not capture the ceteris paribus effect of X_1 on Y (as a change in X_1 leads to a change in X_2)

The bias in estimating β_1 when X_2 is incorrectly omitted can be summarized in the following Table 6.1:

Table 6.1

	$C(X_1, X_2) > 0$	$C(X_1, X_2) < 0$
$\beta_2 > 0$	+	–
$\beta_2 < 0$	–	+

(b) Including an Irrelevant Variable

When X_2 is considered a variable that is irrelevant, i.e., $\beta_2 = 0$

Then, the real model will be:

$$Y = \gamma_0 + \gamma_1 X_1 + \varepsilon_1$$

When X_2 is considered relevant, it is estimated that:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_1$$

Based on this, as $\beta_2 = 0$ in the true model:

$$\varepsilon = \varepsilon_1 - \beta_2 X_2 = \varepsilon_1$$

This implies that:

$$\begin{aligned} E[\varepsilon \mid X_1, X_2] &= E[\varepsilon_1 - \beta_2 X_2 \mid X_1, X_2] \\ &= E[\varepsilon_1 \mid X_1] \end{aligned}$$

Properties of the OLS Estimators

Based on the above, $\hat{\beta}_1$ is an unbiased and consistent estimator of γ_1, with lower variance than $V(\hat{\beta}_1) \geq V(\hat{\gamma}_1)$, which also is consistent and unbiased.

In other words, if irrelevant variables get included in the analysis, the consistency of the estimated effect of the variables is not affected.

Intuition: The true population value of an irrelevant variable's coefficient is 0. Thus, with this variable's inclusion, there is no effect on the coefficient estimators for the other variables in the limit.

Nevertheless, estimated β's will be generally inefficient. In other words, the variances that they have will be greater than what the true model has.

Intuition: As the correlation between the relevant and irrelevant variables becomes higher and higher, there is an increase in the variance of the estimated coefficient for the relevant variables. In other words, when there is inclusion of irrelevant variables, the problem of inconsistency of the OLS estimator does not exist.

Nevertheless, there can be a serious consequence created by the inefficiency problem when testing hypotheses of type $H_0 : \beta_j = 0$. This is because of the loss of power. Inference can be drawn to say that they are not relevant variables, when they actually are (type II error).

It is not possible to decide which model is appropriate and when to put it to practical use.

Thus, the most appropriate way to proceed is by including only those variables which, based on economic theory, will affect the dependent variable, and are not accounted for other variables in the model. This will enable the collecting of evidence against or for the 'irrelevance' or the 'relevance' of the variable(s) by testing hypotheses.

Let us understand this with the help of an example:

Example 6.1:

- There are 2 groups: non-smokers and smokers
- Information is gathered on cases of cancer amongst the groups.
- There is greater likelihood of smokers engaging in physical activities that leads to a decreased likelihood of cancer. It is not possible for the study to observe such activities.

This gives rise to the possibility of there being an overestimation of the impact of smoking on cancer due to the fact that tobacco consumption could reduce physical activity levels.

This all can be depicted as:

$C_i = \beta_0 + \beta_1 F_i + \beta 2_E J_i + \varepsilon,$

In the above equation:

C_i represents the measure of cancer for individual i

F_i is a dummy variable that assumes value 1 for a smoker and 0 for non-smoker

EJ_i represents the measure of physical activity (exercise)

The true values are considered to be:

$\beta_1 > 0, \beta_2 < 0$

Furthermore,

$$C_i = \delta_0 + \delta_1 F_i + \varepsilon_1$$

Where,

$\delta_1 < 0$

So, when a simple regression is performed of C_i on F_i, it provides:

NOTES

$$C_i = \gamma_0 + \gamma_1 F_i + \varepsilon_2$$

This further provides:

$$\hat{\gamma}_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1$$

Let us take this example further.

Besides health impacts, there are also economic consequences of smoking.

For example, wages of smokers might be lower compared with non-smokers. The reasons for this could be their lower productivity because of taking cigarette breaks, the greater probability of being sick and being on leave, and so on.

Here is actual representative data from USA for 30 years old individuals. A wage equation was estimated by Levine, Gustafson and Velenchik (1997) and they used the variables as given below:

Y = ln (wage)

F = a dummy variable that takes a value of 1 for smokers and 0 for non-smokers

ED = Years of education

It has to be considered that on an average, non-smokers are more educated than smokers. So, there is a negative correlation of education with smoking.

There are 2 specifications being considered:

Education is omitted:

$$\hat{Y}_i = -0.176F_i \text{ with } s_{\hat{\beta}_1} = 0.021$$

Education is included:

$$\hat{Y}_i = -0.080F_i + 0.070ED_i$$

$$\text{with: } s_{\hat{\beta}_1} = 0.021 \text{ and } s_{\hat{\beta}_2} = 0.004$$

If education is not included in the regression, the impact of smoking will be overestimated.

(c) Error of Measurement

The assumption has been till now that Y and X_j are without errors of measurement and therefore accurate.

It has to be remembered that at times there will be data with measurement errors. There can also be times that there is a lack of availability of data for the variable under consideration.

Consider the following example. The life cycle models consider that consumption is dependent upon permanent income that cannot be measured without error, or, we have data on the reported annual income, but not the real annual income.

A measurement error will happen in case accurate measurement cannot be made of the variable of interest's magnitude. The point of interest is how such errors affect the OLS estimators.

Measurement error can be of two types:

- Those that occur in the independent variable, X
- Those that occur in the dependent variable, Y

For measurement errors in the dependent variable, Y, consider the following model.

$$Y^* = \beta_0 + \beta_1 X + \varepsilon \tag{6.5}$$

$$E[\varepsilon \mid X] = 0 \Rightarrow E[Y^* \mid X] = L(Y^* \mid X) = \beta_0 + \beta_1 X$$

$$E(\varepsilon) = 0, C(X, \varepsilon) = 0 \text{ and } \beta_0 = E[Y^*] - \beta_1 E[X],$$

$$\beta_1 = \frac{C(X, Y^*)}{V(X)}$$

Consider that there is a measurement error in the Equation (6.5), *Y**, such that:

$$Y = Y^* + \upsilon_0$$

In which, υ_0 is the error of measurement in *Y**.

In this case the model will be:

$Y = \beta_0 + \beta_1 X + \varepsilon + \upsilon_0$, which gives

$$Y = \beta_0 + \beta_1 X + \underbrace{(\varepsilon + \upsilon_0)}_{u} \tag{6.6}$$

Here, *u* is the composite error term.

As was shown before:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_i x_i y_i}{\frac{1}{n}\sum_i x_i^2}, y_i = Y_i - \bar{Y}, \text{ and } x_i = X_i - \bar{X}$$

$$p_{n\to\infty}^{\lim}\hat{\beta}_1 = \frac{p_{n\to\infty}^{\lim}\frac{1}{n}\sum_i x_i y_i}{p_{n\to\infty}^{\lim}\frac{1}{n}\sum_i x_i^2} = \frac{C(X, Y)}{V(X)}$$

$$= \frac{C(X, Y^* + \upsilon_0)}{V(X)} = \frac{C(X, Y^*) + C(X, \upsilon_0)}{V(X)}$$

NOTES

$$= \beta_1 + \frac{C(X, \upsilon_0)}{V(X)} \Rightarrow \text{ consistent if } C(X, \upsilon_0) = 0$$

In case of $\hat{\beta}_0$

$$p^{lim}_{n\to\infty}\hat{\beta}_0 = p^{lim}_{n\to\infty}\left(\bar{Y} - \hat{\beta}_1\bar{X}\right)$$

$$= E\left[Y^* + \upsilon_0\right] - E[X] p^{lim}_{n\to\infty}\hat{\beta}_1$$

Thus, $\hat{\beta}_0$ and $\hat{\beta}_1$ are consistent when,

- $C(X, \upsilon_0) = 0$

and

- $E[\upsilon_0] = 0$

Now, the variances will be larger under measurement error:

- $V(Y^* \mid X) = V(\varepsilon \mid X) = \sigma^2$

  $V(V|X) = V(Y^* + \upsilon_0 \mid X) = \sigma^2 + \sigma^2_{\upsilon_0}$, having assumed that $C(\varepsilon, \upsilon_0|X) = 0$

Measurement Error in X

Look at the model given below:

$$Y = \beta_0 + \beta_1 X^* + \varepsilon \qquad (6.7)$$

$E[\varepsilon \mid X^*] = 0 \Rightarrow E[Y \mid X^*] = L(Y \mid X^*) = \beta_0 + \beta_1 X^*$

$E(\varepsilon) = 0$, $C(X^*, \varepsilon) = 0$ and $\beta_0 = E[Y] - \beta_1 E[X^*]$,

$$\beta_1 = \frac{C(X^*, Y)}{V(X^*)}$$

Consider that in Equation (6.7) there is measurement error in *X**, such that:

$X = X^* + \upsilon_1$

with υ_1 being the measurement error.

The model will then be:

$$Y = \beta_0 + \beta_1(X - \upsilon_1) + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \underbrace{(\varepsilon - \beta_1\upsilon_1)}_{\upsilon_1} \qquad (6.8)$$

There is no correlation of

$E[\upsilon_1] = 0$, and $C(X, \varepsilon) = 0$; ε with X, X^* or υ_1. $E[Y \mid X^*, X] = E[Y \mid X^*] \Rightarrow$ Given X^*. There is no additional relevant information contained in X.

NOTES

Assuming that the classical measurement error assumptions are true. That is to say:

$$C(X^*, v_1) = 0 \text{ and } C(v_1, \varepsilon) = 0$$

$$\begin{aligned}
p \lim_{n\to\infty} \hat{\beta}_1 &= \frac{p \lim\limits_{n\to\infty} \frac{1}{n} \sum_i x_i y_i}{p \lim\limits_{n\to\infty} \frac{1}{n} \sum_i x_i^2} = \frac{C(X,Y)}{V(X)} \\
&= \frac{C(X^* - v_1, Y)}{V(X^* - v_1)} = \frac{C(X^*, Y) + \overbrace{C(Y, v_0)}^{0}}{V(X^*) + V(v_1)} \\
&= \frac{C(X^*, Y)/V(X^*)}{(V(X^*) + V(v_1))/V(X^*)} \\
&= \frac{\beta_1}{1 + V(v_1)/V(X^*)} \neq \beta_1
\end{aligned}$$

Then, the asymptotic bias is:

$$\begin{aligned}
p \lim_{n\to\infty} \left(\hat{\beta}_1 = \beta_1\right) &= \frac{\beta_1}{1 + V(v_1)/V(X^*)} - \beta_1 \\
&= -\beta_1 \frac{V(v_1)}{V(v_1) + V(X^*)}
\end{aligned}$$

In the case of a multiple regression model, the overall measurement error in an explanatory variable will lead to an inconsistency of each one of the estimators, meaning that all $\hat{\beta}_j$'s will be inconsistent. For a multiple regression model with measurement error in just a single X_j and with the error not being correlated with either wrongly measured, say X_m or with the remaining X_j's $(m \neq j)$:

- There will be bias in $\hat{\beta}_m$ and it will also be inconsistent.
- In case of the rest of the $\hat{\beta}_j$'s $(m \neq j)$, they will generally be inconsistent, though the inconsistency and bias magnitude and direction is difficult to know.
- Only if X_m is orthogonal with X_j's $(m \neq j)$, (though this is unlikely), will $\hat{\beta}_j$'s be inconsistent.

Let us understand this with an example regarding the effect that family income has on college grade point average.

As there is no clarity regarding whether there is actually any direct effect of family income on academic performance, it is recommended that the strategy to use is including this variable as a regressor and testing if it provides a significant coefficient.

NOTES

Thus,

CAL = $\beta_0 + \beta_1 I^* + \beta_2$PRE + β_3SEL + ε

In the above:

- CAL represents the average grade in college
- I^* depicts the family income,
- PRE depicts the average grade prior college entrance
- SEL depicts the average grade on the admission exam

Suppose the data is obtained by surveying students. There may be errors in the declared family incomes, so $I = I^* + \upsilon_1$.

If we consider the measurement error, υ_1, to not be uncorrelated with I^* or any other explanatory variable (PRE, SEL), the estimators that will be obtained by employing I rather than employing the true value I^* will be inconsistent.

There will be an underestimation of $|\beta_1|$. Therefore, when we test how significant β_1 is, it is most possible that (DNR) H_0 will not be rejected.

The example given above, is one in which it is difficult to assess the direction and magnitude of the bias and the inconsistency for the estimators of β_2 and β_3.

Example 6.2: Let us examine the impact of consumption of tobacco on users.

(i) There is proof of cancer in a group of non-smokers and a group of smokers.

(ii) The smokers are more physically active than the non-smokers, while we cannot observe the activity, and this leads to a reduced risk of cancer.

(iii) The impact that smoking has on cancer could be overestimated due to consumption of tobacco reducing the physical activity level.

Thus

$$C_i = \beta_0 + \beta_1 F_i + \beta 2_E J_i + \varepsilon$$

Where:

C_i as the measure of cancer for individual i

F_i is a dummy variable that takes a value of 1 is the individual i is a smoker and 0 otherwise

EJ_i is a measure of physical activity, i.e., exercise

Let the true values be $\beta_1 > 0$, $\beta_2 < 0$

In addition,

$$C_i = \delta_0 + \delta_1 F_i + \varepsilon_1, \text{with } \delta_1 < 0$$

Thus, running the simple regression of C_i on F_i, we get:

$$C_i = \gamma_0 + \gamma_1 F_i + \varepsilon_2 ,$$

Then we have $\hat{\gamma}_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1$

Besides impacting health, does smoking have an economic impact?

The answer is yes. Non-smokers might have higher wages than smokers if:

- Non-smokers are less productive due to 'cigarette breaks'
- Smoking has an impact on health outcomes, smokers would more likely ask for sick-leaves
- The organisation discriminates against smokers; and such other reasons.

NOTES

On average, smokers are less educated than non-smokers in Western countries, making education negatively correlated with smoking.

Two of the specifications that have been taken into account are:

Omitting education: $\hat{Y}_i = -0.176F_i$ with $s_{\hat{\beta}_1} = 0.021$

Including education: $\hat{Y}_i = -0.080F_i + 0.070ED_i$ with $s_{\hat{\beta}_1} = 0.021$ and $s_{\hat{\beta}_2} = 0.004$

Excluding education from the regression leads to an overestimation of the impact that smoking can have.

Check Your Progress

1. Explain the basic concept of specification analysis.
2. Define the omission of a relevant variable.
3. Illustrate the inclusion of irrelevant variable.
4. Elaborate on the test of specification errors.
5. What do you understand by the error of measurement?

6.6 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. In econometrics, specification tests plays an important role to verify the validity of one specification at a time. It is said that most of these tests are not, in general, strong in the presence of other misspecifications. These tests will 'Confirm' the validity (or invalidity) of a general model requiring the estimates of the restricted model only.
2. The analyst may delete some of the explanatory variables to keep the model simple. There can be numerous causes behind such decisions, e.g., it may be hard to quantify the variables like the taste, intelligence, etc. Sometimes, it may be difficult to take correct observations on the variables like income etc.

NOTES

3. Some variables may contribute very little to the explanatory power of the model. This may tend to reduce the degrees of freedom ($n–k$), and consequently, the validity of inference drawn may be questionable. For example, the value of the coefficient of determination will increase, indicating that the model is getting better, which may not really be true.
4. In a linear regression model, the process of model specification requires three key decisions:
   - o Selection of the independent variables,
   - o Omitting of the variables,
   - o Selecting the functional form.
5. The assumption has been till now that Y and X_j are without errors of measurement and therefore accurate. It has to be remembered that at times there will be data with measurement errors. There can also be times that there is a lack of availability of data for the variable under consideration.

6.7 SUMMARY

- In econometrics, specification tests plays an important role to verify the validity of one specification at a time. It is said that most of these tests are not, in general, strong in the presence of other misspecifications. These tests will 'Confirm' the validity (or invalidity) of a general model requiring the estimates of the restricted model only.
- Model specification is part of the process of building a statistical model: specification consists of selecting an appropriate functional form for the model and choosing which variables to include.
- Specification error occurs when the functional form or the choice of independent variables poorly represent relevant aspects of the true data-generating process. In particular, bias (the expected value of the difference of an estimated parameter and the true underlying value) occurs if an independent variable is correlated with the errors inherent in the underlying process.
- The Durbin–Wu–Hausman test (also called Hausman specification test) is a statistical hypothesis test in econometrics named after James Durbin, De-Min Wu, and Jerry A. Hausman. The test evaluates the consistency of an estimator when compared to an alternative, less efficient estimator which is already known to be consistent. It helps one evaluate if a statistical model corresponds to the data.
- The Ramsey Regression Equation Specification Error Test (RESET) test is a general specification test for the linear regression model. More specifically, it tests whether non-linear combinations of the fitted values help explain the response variable.

- Common criteria for comparing models include the following: R2, Bayes factor, and the likelihood-ratio test together with its generalization relative likelihood. For more on this topic, see statistical model selection.
- The analyst may delete some of the explanatory variables to keep the model simple. There can be numerous causes behind such decisions, e.g., it may be hard to quantify the variables like the taste, intelligence, etc. Sometimes, it may be difficult to take correct observations on the variables like income etc.
- Some variables may contribute very little to the explanatory power of the model. This may tend to reduce the degrees of freedom ($n–k$), and consequently, the validity of inference drawn may be questionable.
- In a linear regression model, the process of model specification requires three key decisions:
  - Selection of the independent variables,
  - Omitting of the variables,
  - Selecting the functional form.
- The assumption has been till now that Y and X_j are without errors of measurement and therefore accurate. It has to be remembered that at times there will be data with measurement errors. There can also be times that there is a lack of availability of data for the variable under consideration.

6.8 KEY WORDS

- **Specification tests:** In econometrics, specification tests plays an important role to verify the validity of one specification at a time. It is said that most of these tests are not, in general, strong in the presence of other misspecifications.
- **Omission of a relevant variable:** The analyst may delete some of the explanatory variables to keep the model simple. There can be numerous causes behind such decisions, e.g., it may be hard to quantify the variables like the taste, intelligence, etc.
- **Inclusion of irrelevant variable:** Some variables may contribute very little to the explanatory power of the model. This may tend to reduce the degrees of freedom ($n–k$), and consequently, the validity of inference drawn may be questionable.
- **Error of measurement:** The assumption has been till now that Y and X_j are without errors of measurement and therefore accurate. It has to be remembered that at times there will be data with measurement errors.

6.9 SELF ASSESSMENT QUESTIONS AND EXERCISES

NOTES

Short-Answer Questions

1. Define the basic concept of specification analysis.
2. Explain the omission of a relevant variable.
3. Interpret the inclusion of irrelevant variable.
4. What do you understand by the test of specification errors?
5. Elaborate on the error of measurement.

Long-Answer Questions

1. Discuss briefly the basic concept of specification analysis. Give appropriate examples.
2. Explain the omission of a relevant variable. Write the properties of the OLS estimators for this.
3. Describe the inclusion of irrelevant variable with the help of examples. Give the properties of the OLS estimators for this.
4. Analyse the test of specification errors.
5. Define the error of measurement.

6.10 FURTHER READINGS

Johnston, J. and John DiNARDO. 1997. *Econometric Methods*, Fourth Edition. New Delhi: Tata McGraw-Hill.

Koutsoyiannis, A.1977. *Theory of Econometrics,* Second Edition. London: The Macmillan Press Ltd.

Özdemir, Durmuº. 2016. *Applied Statistics for Economics and Business,* Second Edition. Izmir (Turkey): Springer.

Maddala, G. S. 1992. *Introduction to Econometrics,* Second Edition. New York: Macmillan Publishing Company.

Pindyck, R. S and D. L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts,* Fourth Edition. New York: McGraw Hill.

Goldberger, A. S. 1998. *Introductory Econometrics***.** Cambridge: Harvard University Press.

Levine, David M., Timothy C. Krehbiei, Mark L. Berenson and P. K. Viswanathan. 2009. *Business Statistics*, Fifth Edition. New Delhi: Pearson Education.

Webster, Allen L. 1998. *Applied Statistics for Business and Economics,* Third Edition. New Delhi: Tata McGraw-Hill.

UNIT 7 PANEL DATA MODELS

NOTES

Structure

7.0 INTRODUCTION

In econometrics, panel data and longitudinal data are both multi-dimensional data involving measurements over time. Panel data is a subset of longitudinal data where observations are for the same subjects each time.

Time series and cross-sectional data can be thought of as special cases of panel data that are in one dimension only (one panel member or individual for the former, one time point for the latter). A study that uses panel data is called a longitudinal study or panel study.

A balanced panel is a dataset in which each panel member (i.e., person) is observed every year. Consequently, if a balanced panel contains N panel members and T periods, the number of observations (n) in the dataset is necessarily $n = N \times T$. An unbalanced panel is a dataset in which at least one panel member is not observed every period. Therefore, if an unbalanced panel contains N panel members and T periods, then the following strict inequality holds for the number of observations (n) in the dataset: $n < N \times T$.

Panel (data) analysis is a statistical method, widely used in social science, epidemiology, and econometrics to analyse two-dimensional (typically cross sectional and longitudinal) panel data. The data are usually collected over time and over the same individuals and then a regression is run over these two dimensions. Multidimensional analysis is an econometric method in which data are collected over more than two dimensions (typically, time, individuals, and some third dimension).

NOTES

There are unique attributes of individuals that do not vary over time. That is, the unique attributes for a given individual *i* and *t* invariant. These attributes may or may not be correlated with the individual dependent variables y_i. To test whether fixed effects, rather than random effects, is needed, the Durbin–Wu–Hausman test can be used.

There are unique, time constant attributes of individuals that are not correlated with the individual regressors. Pooled OLS{clarify|What is pooled OLS?|date=June 2021}} can be used to derive unbiased and consistent estimates of parameters even when time constant attributes are present, but random effects will be more efficient. Fixed effects is a feasible generalised least squares technique which is asymptotically more efficient than Pooled OLS when time constant attributes are present. Random effects adjusts for the serial correlation which is induced by unobserved time constant attributes.

In this unit, you will study about the panel data models, methods of estimation, fixed effects model, and random effects model.

7.1 OBJECTIVES

After going through this unit, you will be able to:

- Elaborate on the panel data models
- Explain the methods of estimation
- Define the fixed effects model
- Analyse the random effects model

7.2 PANEL DATA MODELS

Econometric Estimations can be conducted on varied data set. Three types of data are generally available for empirical analysis like time series, cross section and panel data estimation. In time series the value of one or more variable is observed over a period of time (GDP for several years). In cross section data is collected on one or more variables are collected for various sample units at the same point in time (number of girl child per house hold for 29 states of India for a given year). In panel data estimation the same cross sectional unit is captured over the period of time. It can be stated that panel data have space as well as time dimension. Panel data can also be referred to as pooled data (created using time series and cross sectional data points), longitudinal data and event history analysis.

Panel data consist of recurring observations over the period of time on the same set of cross-sectional observations. These observations can be individuals, firms, schools, cities, or set of observations one can survey over time. Special econometric methods have been planned to diagnose and exploit the rich evidence available in panel data sets. Time dimension is a key feature of panel data

observations, issues of multicollinearity and dynamic effects need to be focused upon. Additional, different from cross-sectional data, panel data sets allow the presence of systematic, unobserved differences across units.

NOTES

7.2.1 Advantages of Panel Data Estimation

1. Panel data can take categorical explanation of individual-specific heterogeneity ("Individual" here means related to the micro units like firms, cities, country etc.).
2. Panel data by combining data in two dimensions gives more data variation, reduced multicollinearity and allows more degrees of freedom.
3. Panel data is well equipped in comparison to cross-sectional data for understanding the *dynamics of change*. For example, it is well suited to understanding *evolution* behaviour – for example, company failure or growth through merger and acquisitions.
4. Panel data is improvisation over cross section or time series data for better at detecting and measuring effects of the model.
5. Panel data enables the study of more complex behavioural models. For example, the effects of innovations, or business cycles.
6. Panel data reduces the effects of aggregation bias and clearly identifies difference between individual and group effects.

7.2.2 Balanced and Unbalanced Panel

If each cross-sectional unit has the same number of time series observations, then such a panel (data) is called a balanced panel. If the number of observations differs among panel units or have different number of time series observations, such a panel is called as an unbalanced panel.

A matrix of balanced panel data observations on variable y_i, N cross-sectional observations, T time series observations.

$$\begin{bmatrix} y_{11} & y_{21} & \cdots & y_{i1} & \cdots & y_{N1} \\ y_{12} & y_{22} & \cdots & y_{i2} & \cdots & y_{N2} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ y_{1t} & y_{2t} & \cdots & y_{it} & \cdots & y_{Nt} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ y_{1T} & y_{2T} & \cdots & y_{iT} & \cdots & y_{NT} \end{bmatrix}$$

Panel Data: A Well-Known Example

To set the stage, a concrete example is discussed below. Consider the model given which is taken from a famous study of investment theory proposed by Y. Grunfeld.

NOTES

Grunfeld examined how real gross investment (Y) depends on the real value of the firm (X2) and real capital stock (X3)? Although, the original study covered several companies for sake of simplicity the data on four companies, General Electric (GE), General Motor (GM), U.S. Steel (US), and Westinghouse is examined. Data for each company on the above mentioned three variables are available for the period 1935–1954. Hence, there are four cross-sectional variables and 20 time periods. In all, therefore, we have 80 observations. By theory, Y is expected to be positively related to X2 and X3. By standard way we could run four time series regressions, one for each company or we could run 20 cross-sectional regressions, one for each year, causing problems with degrees of freedom.

Pooling, or combining, all the 80 observations, the Grunfeld investment function can be written as:

$$Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X3_{it} + u_{it} \quad (7.1)$$

$i = 1, 2, 3, 4$

$t = 1, 2, \ldots, 20$

Where i represents the i-th cross-sectional unit and t represents the t-th time period. As convention, let i denote the cross-section identifier and t the time identifier.

7.3 ESTIMATION OF PANEL DATA REGRESSION MODELS

Panel data models can be estimated using fixed effect and random effect.

7.3.1 Fixed Effect Estimation Approach

Fixed effects studies variables that are constant across individuals. In a fixed effects model, the unobserved variables are permitted to have any relations whatsoever with the observed variables. Fixed effects models control for, or restrict out, the effects of time-invariant variables with time-invariant effects.

Estimates of fixed effects depend on the assumptions made on intercept, slope and the error term u_{it}.

Below are the assumptions for several model:

1. The intercept and slope coefficients are constant across time and space.
2. The slope coefficients are constant but the intercept varies over individuals.
3. The slope coefficients are constant but the intercept varies over individuals and time.
4. All coefficients (the intercept as well as slope coefficients) vary over individuals.

Using the example of Grunfeld investment theory the discussion will capture each possibility of fixed effect:

A. All Coefficients slope and intercept Constant across Time and Individuals: The easiest, approach is to ignore the space and time dimensions of the pooled data and simply estimate the usual OLS regression. That is, pile the 20 observations for each company one on top of the other, creating in all 80 observations for each of the variables in the model. The OLS results are as follows:

$$\hat{Y} = -63.3041 + 0.1101X2 + 0.3034X3 \quad (7.2)$$

se = (29.6124) (0.0137) (0.0493)

t = (–2.1376) (8.0188) (6.1545)

$R^2 = 0.7565$

Durbin–Watson = 0.2187

N (observations) = 80

Degrees of freedom = 77

Where Y is real gross investment, X_2 is real value of firm and X_3 is real stock capital.

We observe that all the coefficients are individually statistically significant, the slope coefficients have the estimated positive signs and the R^2 value is high and acceptable. As anticipated, Y is positively related to X2 and X3. The estimated Durbin–Watson statistic is pretty low suggesting the issue of autocorrelation or due to specification error assuming same intercept value to all four firms and slope coefficients identical for all four firms .The pooled regression in Equation 7.2 may alter the true picture of the relationship between Y and the X's across the four companies. Hence, it is important to take into account the specific nature of the four companies.

B. Slope Coefficients Constant but the Intercept Varies across Individuals: The Fixed Effects or Least-Squares Dummy Variable (LSDV) Regression Model: To take into account the "Individuality" of each cross-sectional unit, i.e., the company in this example is to allow the intercept vary for each company but still assumes that the slope coefficients are constant across companies. To see this, write model equation 1 as:

$$Y_{it} = \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \quad (7.3)$$

Notice subscript I on the intercept term to suggest that the intercepts of the four companies may be different; the differences may be due to uniqueness of each company, such as management or employees work ethics.

As convention, model (7.2) is known as the Fixed Effects (regression) Model (FEM). The term "Fixed Effects" because, though the intercept may differ across individuals (here the four companies), each individual's intercept is fixed over time; that is, it is time invariant. Notice that to write the intercept as β_{1i} , it will suggest

NOTES

that the intercept of each company or individual is time invariant. It may be noted that the FEM undertakes that the slope coefficients of the regressors do not depend on time or individual.

The (fixed effect) intercept can vary between companies by the dummy variable technique particularly, the differential intercept dummies.

Therefore, we write (Equation 7.3) as:

$$Y_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \quad (7.4)$$

Where D_{2i} = 1 if the observation belongs to General Motors, 0 otherwise; D_{3i} = 1 if the observation belongs to US Steels, 0 otherwise; and D_{4i} = 1 if the observation belongs to Westinghouse, 0 otherwise. Three dummies are to be used for four companies to avoid falling into the dummy-variable trap (i.e., the situation of perfect collinearity). Here, there is no dummy for GE. Hence, it can be said, α1 represents the intercept of GE and α2, α3, and α4, the differential intercept coefficients, explaining by how much the intercepts of GM, US, and WEST differ from the intercept of GE. All companies are compared to GE. Since dummies are used to estimate the fixed effects, the model is called as is also known as the Least-Squares Dummy Variable (LSDV) model.

The results based on (Equation 7.4) are as follows:

$$\hat{Y}_{it} = -245.7924 + 161.5722D2i + 339.6328D3i + 186.5666D4i + 0.1079X2i + 0.3461X3i \quad (7.5)$$

se = (35.8112) (46.4563) (23.9863) (31.5068) (0.0175) (0.0266)

t = (“6.8635) (3.4779) (14.1594) (5.9214) (6.1653) (12.9821)

R^2 = 0.8345 Durbin Watson = 1.1076 degrees of freedom = 74

Compare this regression with Equation 7.1. In Equation 7.5 all the estimated coefficients are individually highly significant as depicted by *t* values. The intercept values of the four companies are statistically different; being –245.7924 for GE, –84.220 (=–245.7924 + 161.5722) for GM, 93.8774 (=–245.7924 + 339.6328) for US, and –59.2258 (=–245.7924 + 186.5666) for WEST. These differences in the intercepts reflect the specific characteristics of each company as employee talent or managerial prowess.

By the statistical significance of the estimated coefficients, the R2 value and the Durbin Watson value is higher, suggesting that model of Equation 7.5 in comparison to Equation 7.2 suggests that model mentioned in equation 1 is ill-specified.

Analysing Time Effect

The dummy variables are used to account for individual (company) effect, similarly dummy can be introduced for time effect in the sense that the Grunfeld investment

function shifts over time because of factors such as technological advancement, fiscal policies or any other external factors. The time effects in the model can be easily accounted for if we introduce time dummies, one for each year. Since the data in the study was available for 20 years, from 1935 to 195419 dummies can be introduced and model can be expressed as:

NOTES

$$Y_{it} = \lambda_0 + \lambda_1 D_{um35} + \lambda_2 D_{um36} + \cdots + \lambda_1 9 D_{um53} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \quad (7.6)$$

Where Dum35 takes a value of 1 for observation in year 1935 and 0 otherwise, etc.

C. Constant Slope Coefficients and Intercept Varies over Individuals As Well As Time: To understand the same combine Equations 7.5 and 7.6, as follows:

$$Y_{it} = \alpha_1 + \alpha_2 D_{GMi} + \alpha_3 D_{USi} + \alpha_4 D_{WESTi} + \lambda_0 + \lambda_1 D_{um35t} + \cdots + \lambda_{19} D_{um53t} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \quad (7.7)$$

Regression results provide company dummies as well as the coefficients of the X individually statistically significant, but none of the time dummies are significant. The overall inference that appears is that perhaps there is distinct individual company effect but no time effect. In other words, the investment functions for the four companies are the same except for their intercepts. Hence proved, the X variables had a strong impact on Y but time has no significant impact.

D. All Coefficients Vary across Individuals: The model frames that the intercepts and the slope coefficients are different for all individual, or cross-section, units. This is to say that the investment functions of GE, GM, US, and WEST are all different. LSDV model is applied in such context by introducing individual dummies in an additive manner. However, interactive, or differential, slope dummies, can account for differences in slope coefficients. In the context the Grunfeld investment function, multiply each of the company dummies by each of the X variables and estimate the model.

$$Y_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \beta_2 X_{2it} + \beta_3 X_{3it} + \gamma_1 (D_{2i} X_{2it}) + \gamma_2 (D_{2i} X_{3it}) + \gamma_3 (D_{3i} X_{2it}) + \gamma_4 (D_{3i} X_{3it}) + \gamma_5 (D_{4i} X_{2it}) + \gamma_6 (D_{4i} X_{3it}) + u_{it} \quad (7.8)$$

γ 's are the differential slope coefficients, just as α_2, α_3, and α_4 are the differential intercepts. If one or more of the γ coefficients are statistically significant, it explains that slope coefficients are different from the base individual. If, β_2 and γ_1 are statistically significant. In this case $(\beta_2 + \gamma_1)$ will give the value of the slope coefficient of X_2 for General Motors and the same is different from other companies.

If all the differential intercept and all the differential slope coefficients are statistically significant, the conclusion is that the investment functions of General Motors, United States Steel, and Westinghouse are different from that of General Electric. Hence the estimate is complete improvisation on equation 7.2.

Table 7.1 *Results of Regression using Grunfeld Data*

Variable	Coefficient	Std. error	*t* value	*p* value
Intercept	-9.9563	76.3518	-0.1304	0.8966
D_{2i}	-139.5104	109.2808	-1.2766	0.2061
D_{3i}	-40.1217	129.2343	-0.3104	0.7572
D_{4i}	9.3759	93.1172	0.1006	0.9201
X_{2i}	0.0926	0.0424	2.1844	0.0324
X_{3i}	0.1516	0.0625	2.4250	0.0180
$D_{2i}X_{2i}$	0.0926	0.0424	2.1844	0.0324
$D_{2i}X_{3i}$	0.2198	0.0682	3.2190	0.0020
$D_{3i}X_{2i}$	0.1448	0.0646	2.2409	0.0283
$D_{3i}X_{3i}$	0.2570	0.1204	2.1333	0.0365
$D_{4i}X_{2i}$	0.0265	0.1114	0.2384	0.8122
$D_{4i}X_{3i}$	-0.0600	0.3785	-0.1584	0.8745
	$R^2 = 0.9511$	$d = 1.0896$		

NOTES

Interpreting the regression results mentioned in below table based on Equation 7.8 reveals that Y is significantly related to X2 and X3. However, several differential slope coefficients are statistically significant. Notably none of the differential intercepts are statistically significant indicating that investment function of four companies is different from each other.

7.3.2 The Random Effect Model

In a random effects model, the unobserved variables are considered to be uncorrelated with all the observed variables. Random Effect allows estimation for variables. Using dummy variable may not express true model in fixed effect estimation. This concern can be addressed through the disturbance term u_{it}. Random Effect model also known as error component method uses this approach.

The basic idea is to start with Equation 7.3

$$Y_{it} = \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \quad (7.9)$$

In place of considering β_{1i} as fixed, It is assumed that it is a random variable with a mean value of β_1 (no subscript *i* here). And the intercept value for each company can be expressed as:

$$\beta_{1i} = \beta_1 + \varepsilon_i \quad (7.10)$$

$$i = 1, 2, \ldots, N$$

Where ε_i is a random error term with a mean zero and variance of σ^2_ε.

The example showing four companies express that the sample is drawn from large population of companies and have common mean for the intercept = β_1 and the individual differences in the intercept of each company are expressed in the error term ε_i.

Substituting Equation 7.10 into Equation 7.9, we get

$$Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + \varepsilon_i + u_{it}$$

$$= \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + w_{it} \quad (7.11)$$

Where

$w_{it} = \varepsilon_i + u_{it}$

The combined error term w_{it} consists of two components, ε_i, which is the cross-section, or individual error component, and u_{it}, which is the pooled error component. The usual assumptions made by random effect model are:

NOTES

$\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$

$u_{it} \sim N(0, \sigma^2_u)$ (7.12)

$E(\varepsilon_i, u_{it}) = 0$ $E(\varepsilon_i \varepsilon_j) = 0$ (i is not equal to j)

$E(u_{it}u_{is}) = E(u_{it}u_{jt}) = E(u_{it}u_{js}) = 0$ (i is not equal to j ; t is not equal to s).

The above expressions explains that , the individual error components are not correlated with each other and are not auto correlated across both cross-section and time series data points.

The results of Random Effect estimation model for the Grunfeld investment function are given in Table 7.2. Important to note is that the summation of the random effect values given for the four companies will be zero. The mean value of the random error component, ε_i, is the common intercept value of –73.0353. The random effect value of GE of –169.9282 explains by how much the random error component of GE differs from the common intercept value and same follows for other companies.

Table 7.2 *ECM Estimation of the Grunfeld Investment Function*

Variable	Coefficient	Std. error	t statistic	p value
Intercept	-73.0353	83.9495	-0.8699	0.3870
X_2	0.1076	0.0168	6.4016	0.0000
X_3	0.3457	0.0168	13.0235	0.0000
Random effect:				
GE	-169.9282			
GM	-9.5078			
USS	165.5613			
Westinghouse	13.87475			
	R^2 = 0.9323 (GLS)			

7.3.3 Choosing between Fixed Effects (FE) and Random Effects (RE) Models

Researchers face challenges that which model to be chosen Fixed Effect or Random Effect. The answer lies in the assumption the researcher make about the likely correlation between the cross section specific error component and regressors X. However, researcher can focus on below points for selecting the model:

1. In case of large time series data and small cross section observations there is likely to be little difference between two models hence, FE is preferable and easier to compute against Random effect.

NOTES

2. With large cross section units and small time series observations, estimated parameters can differ significantly. In case the cross-sectional groups are a random sample of the population Random Effect is preferable otherwise fixed effect to be chosen?
3. If the error component is correlated with regressors then RE parameters will be biased, but FE is not. Hence, FE results must be chosen.
4. With large cross section units and small time series observations along with the assumptions behind Random Effect hold then RE estimators is more efficient than FE estimators.

7.4 HAUSMAN TEST

However, researchers always look for a formal or objective test for decision making and hence a test was developed by Hausman in 1978 to choose between fixed effect and random effect model. Hausman test is tests for the statistical significance of the difference between the coefficient estimates obtained by FE and by RE, under then null hypothesis that the RE estimates are efficient and consistent, and FE estimates are inefficient. The test statistic developed by Hausman has an asymptomatic Chi^2 distribution. The test has a Wald test form, and is usually reported in Chi^2 form with k-1 degrees of freedom (k is the number of regressors). If W < critical value then random effects is the preferred estimator over fixed effect estimators.

Check Your Progress

1. What do you understand by the panel data?
2. Define the advantages of panel data estimation.
3. Explain the balanced and unbalanced panel.
4. Illustrate the fixed effect estimation approach.
5. Elaborate on the analysing time effect.
6. Interpret the random effect model.
7. Explain the Hausman test.

7.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Panel data consist of recurring observations over the period of time on the same set of cross-sectional observations. These observations can be individuals, firms, schools, cities, or set of observations one can survey over time. Special econometric methods have been planned to diagnose and exploit the rich evidence available in panel data sets.

NOTES

2. Panel data enables the study of more complex behavioural models. For example, the effects of innovations, or business cycles. Panel data reduces the effects of aggregation bias and clearly identifies difference between individual and group effects.
3. If each cross-sectional unit has the same number of time series observations, then such a panel (data) is called a balanced panel. If the number of observations differs among panel units or have different number of time series observations, such a panel is called as an unbalanced panel.
4. Fixed effects studies variables that are constant across individuals. In a fixed effects model, the unobserved variables are permitted to have any relations whatsoever with the observed variables. Fixed effects models control for, or restrict out, the effects of time-invariant variables with time-invariant effects.
5. The dummy variables are used to account for individual (company) effect, similarly dummy can be introduced for time effect in the sense that the Grunfeld investment function shifts over time because of factors such as technological advancement, fiscal policies or any other external factors.
6. In a random effects model, the unobserved variables are considered to be uncorrelated with all the observed variables. Random Effect allows estimation for variables. Using dummy variable may not express true model in fixed effect estimation. This concern can be addressed through the disturbance term u_{it}. Random Effect model also known as error component method uses this approach.
7. Hausman test is tests for the statistical significance of the difference between the coefficient estimates obtained by FE and by RE, under then null hypothesis that the RE estimates are efficient and consistent, and FE estimates are inefficient. The test statistic developed by Hausman has an asymptomatic Chi^2 distribution.

7.6 SUMMARY

- Econometric Estimations can be conducted on varied data set. Three types of data are generally available for empirical analysis like time series, cross section and panel data estimation. In time series the value of one or more variable is observed over a period of time (GDP for several years).
- In panel data estimation the same cross sectional unit is captured over the period of time. It can be stated that panel data have space as well as time dimension. Panel data can also be referred to as pooled data (created using time series and cross sectional data points), longitudinal data and event history analysis.

NOTES

- Panel data consist of recurring observations over the period of time on the same set of cross-sectional observations. These observations can be individuals, firms, schools, cities, or set of observations one can survey over time. Special econometric methods have been planned to diagnose and exploit the rich evidence available in panel data sets.
- Panel data can take categorical explanation of individual-specific heterogeneity ("Individual" here means related to the micro units like firms, cities, country etc.).
- If each cross-sectional unit has the same number of time series observations, then such a panel (data) is called a balanced panel.
- If the number of observations differs among panel units or have different number of time series observations, such a panel is called as an unbalanced panel.
- Fixed effects studies variables that are constant across individuals. In a fixed effects model, the unobserved variables are permitted to have any relations whatsoever with the observed variables. Fixed effects models control for, or restrict out, the effects of time-invariant variables with time-invariant effects.
- The dummy variables are used to account for individual (company) effect, similarly dummy can be introduced for time effect in the sense that the Grunfeld investment function shifts over time because of factors such as technological advancement, fiscal policies or any other external factors.
- In a random effects model, the unobserved variables are considered to be uncorrelated with all the observed variables. Random Effect allows estimation for variables. Using dummy variable may not express true model in fixed effect estimation.
- Hausman test is tests for the statistical significance of the difference between the coefficient estimates obtained by FE and by RE, under then null hypothesis that the RE estimates are efficient and consistent, and FE estimates are inefficient.

7.7 KEY WORDS

- **Econometric estimations:** Econometric Estimations can be conducted on varied data set. Three types of data are generally available for empirical analysis like time series, cross section and panel data estimation.
- **Balanced panel:** If each cross-sectional unit has the same number of time series observations, then such a panel (data) is called a balanced panel.
- **Unbalanced panel:** If the number of observations differs among panel units or have different number of time series observations, such a panel is called as an unbalanced panel.

- **Fixed effect:** Fixed effects studies variables that are constant across individuals. In a fixed effects model, the unobserved variables are permitted to have any relations whatsoever with the observed variables.
- **Analysing time effect:** The time effects in the model can be easily accounted for if we introduce time dummies, one for each year.
- **Random effects:** In a random effects model, the unobserved variables are considered to be uncorrelated with all the observed variables. Random Effect allows estimation for variables.
- **Hausman test:** Hausman test is tests for the statistical significance of the difference between the coefficient estimates obtained by FE and by RE, under then null hypothesis that the RE estimates are efficient and consistent, and FE estimates are inefficient.

7.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Elaborate on the panel data.
2. State the advantages of panel data estimation.
3. Define the balanced and unbalanced panel.
4. Explain the fixed effect estimation approach.
5. Illustrate the analysing time effect.
6. Interpret the random effect model.
7. What do you understand by the Hausman test?

Long-Answer Questions

1. Discuss briefly the panel data. Write the advantages of panel data estimation. Give appropriate example.
2. Analyse the balanced and unbalanced panel with the help of examples.
3. Explain the Fixed Effects Model (FEM). Since, panel data have both time and space observations, how does FEM estimators capture for both space and time dimensions?
4. What is meant by a Random Effect Model and why it is called as Error Component Model (ECM)? How does it differ from FEM?
5. When to use Random Effect model? And when is FEM suitable to use?

7.9 FURTHER READINGS

NOTES

Johnston, J. and John DiNARDO. 1997. *Econometric Methods*, Fourth Edition. New Delhi: Tata McGraw-Hill.

Koutsoyiannis, A.1977. *Theory of Econometrics,* Second Edition. London: The Macmillan Press Ltd.

Özdemir, Durmuº. 2016. *Applied Statistics for Economics and Business,* Second Edition. Izmir (Turkey): Springer.

Maddala, G. S. 1992. *Introduction to Econometrics,* Second Edition. New York: Macmillan Publishing Company.

Pindyck, R. S and D. L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts,* Fourth Edition. New York: McGraw Hill.

Goldberger, A. S. 1998. *Introductory Econometrics***.** Cambridge: Harvard University Press.

Levine, David M., Timothy C. Krehbiei, Mark L. Berenson and P. K. Viswanathan. 2009. *Business Statistics*, Fifth Edition. New Delhi: Pearson Education.

Webster, Allen L. 1998. *Applied Statistics for Business and Economics,* Third Edition. New Delhi: Tata McGraw-Hill.

UNIT 8 REGRESSION ON DUMMY VARIABLES

NOTES

Structure

8.0 INTRODUCTION

In regression analysis, a dummy variable is a numerical variable which is used to represent a subgroups of the sample. Dummy variables are very useful because they assist us to use a single regression equation to represent multiple groups. Dummy variables are used frequently in time series analysis with regime switching, seasonal analysis and qualitative data applications. A dummy variable or an indicator variable, is a numeric variable that represents categorical data, such as gender, race, and political affiliation, etc.

In econometrics, particularly in regression analysis, a dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. They can be thought of as numeric stand-ins for qualitative facts in a regression model, sorting data into mutually exclusive categories (such as smoker and non-smoker).

A dummy independent variable (also called a dummy explanatory variable) which for some observation has a value of 0 will cause that variable's coefficient to have no role in influencing the dependent variable, while when the dummy takes on a value 1 its coefficient acts to alter the intercept. For example, suppose membership in a group is one of the qualitative variables relevant to a regression. If group membership is arbitrarily assigned the value of 1, then all others would get the value 0. Then, the intercept would be the constant term for non-members but would be the constant term plus the coefficient of the membership dummy in the case of group members.

NOTES

Dummy variables are used frequently in time series analysis with regime switching, seasonal analysis and qualitative data applications. A regression model in which the dependent variable is quantitative in nature but all the explanatory variables are dummies (qualitative in nature) is called an Analysis of Variance (ANOVA) model.

In this unit, you will study about the regression on dummy variables, nature of dummy variables, use of dummy variables in seasonal analysis and in combining time series, and cross sectional data.

8.1 OBJECTIVES

After going through this unit, you will be able to:

- Comprehend the regression on dummy variables
- Explain the nature of dummy variables
- Define the uses of dummy variables in seasonal analysis and in combining time series
- Elaborate on the cross sectional data

8.2 NATURE OF DUMMY VARIABLES

Up until now, the variables that we have used in explaining the endogenous variable have a quantitative nature. However, there are other variables of a qualitative nature that can be important when explaining the behaviour of the endogenous variable, such as sex, race, religion, nationality, and geographical region etc. As is obvious, such qualitative factors are not measurable numerically. The following are other examples of these:

(a) When one wants to examine of the relationship between schooling and earnings and one has both males and females in a sample. One would like to see if the sex of the respondent makes a difference.

(b) When one wants to examining the relationship between income and expenditure in Canada, and the sample include both English and French speaking households. One would like to find out whether the ethnic difference is relevant.

(c) When one has data on the GDP growth rate per capita and foreign aid per capita for several developing countries, of which some are democratic while others are authoritarian, and one would like to examine whether the impact of foreign aid on growth is affected by the type of government.

In all the three examples discussed above, one solution is to run separate regressions for the two categories and see whether the coefficients are different.

One can also run a single regression employing all the observations together, measuring the effect of the qualitative factor with what is called a dummy variable. This has two significant advantages of providing a simple way of testing whether the effect of the qualitative factor is significant, and provided that certain assumptions are valid, making the regression estimates more efficient.

Oftentimes, qualitative factors are found to be binary information, for example such as whether an individual is a female or male, single or not, and so forth. Where qualitative factors are found as dichotomous information, it becomes possible to capture the relevant information through defining a zero-one/binary variable. As stated above, in the field of econometrics, binary variables that are employed for use as regressors are referred to as dummy variables.

Use of dummy variables for capturing the effect of qualitative factors

While defining a dummy variable, it is vital to ascertain to which event will the value one be assigned and to which the value zero will be assigned.

Let us take a look at an example.

In gender, we can do the following assigning:

$$male = \begin{cases} 1 & \text{if the person is a male} \\ 0 & \text{if the person is a female} \end{cases}$$

or

$$female = \begin{cases} 1 & \text{if the person is a female} \\ 0 & \text{if the person is a male} \end{cases}$$

Not that both variables, female and male, hold identical information. Employing the use of zero-one variables for qualitative information capturing may be an arbitrary decision, but it allows a natural interpretation of the parameters. We will incorporate dichotomous information into regression models with the use of a simple model of hourly *wage* determination taken as a function of the years of education (*educ*):

$$wage = \beta_1 + \beta_2 educ + u \tag{8.1}$$

For the purpose of measuring discrimination of gender wage, a dummy variable is introduced as an independent variable for the purpose of gender in the above model to obtain the model given below:

$$wage = \beta_1 + \delta_1 female + \beta_2 educ + u \tag{8.2}$$

The gender attribute has two categories: male and female. In the model, the category included is female, and the omitted male category is used as reference category.

NOTES

For Equation 8.1, the graphic depiction is given below considering $\delta_1 < 0$:

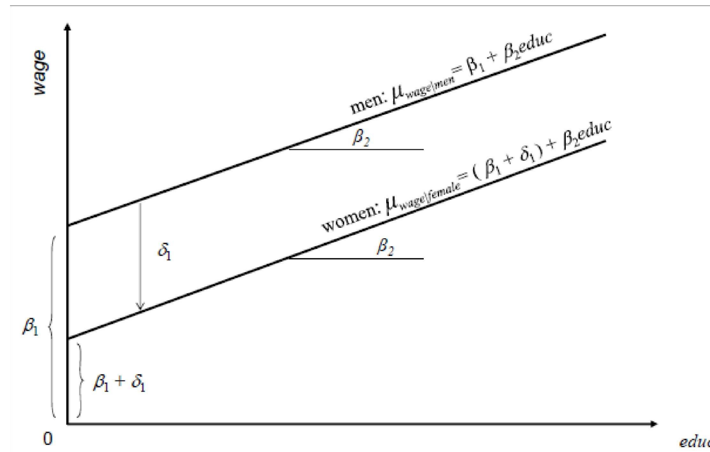


Fig. 8.1

In Figure 8.1, δ_1 represents the difference in hourly wage between males and females based on identical amount of education. So, the coefficient δ_1 determines whether or not there exists any anti-women discrimination. In interpretation, in case of $\delta_1 < 0$, and other factors being the same, then on average, the wages of females are lower than those of males. Considering disturbance mean to be zero, taking expectation for both categories we get:

$$\mu_{wage|female} = E(wage \mid female = 1, educ) = \beta_1 + \delta_1 + \beta_2 educ$$
$$\mu_{wage|male} = E(wage \mid female = 0, educ) = \beta_1 + \beta_2 educ \quad (8.3)$$

In the above equation, for males the intercept is β_1, and for females, the intercept is $\beta_1 + \delta_1$, as had been depicted in the above graphic representation, while there is shift of the intercepts, there are parallel lines for women and men.

In Equation 8.2, for females, a **dummy variable** has been included, but not for males. This has been done as inclusion of dummy variables for both would have caused redundancy. What is actually required are two intercepts (1 for males and 1 for females). When the female dummy variable is introduced, there is an intercept for each gender. If two dummy variables are introduced, there would be perfect multicollinearity since female + male=1, meaning that male is an exact linear function of female as well as of the intercept. The most basic example of this, which is referred to as dummy variable trap, is to include dummy variables for both genders plus the intercept.

In case *male* is used in place of *female*, the wage equation would look as given below:

$$wage = \alpha_1 + \gamma_1 male + \beta_2 educ + u \quad (8.4)$$

The new equation does show have any changed other than how α_1 is interpreted and γ_1: α_1 being the intercept for women, that has become the *reference category* in this equation. Also, the intercept for men is $\alpha_1+\gamma_1$. So, the following is now the relationship between the coefficients:

$$\alpha_1=\beta_1+\delta_1 \text{ and } \alpha_1+\gamma_1=\beta_1 \Rightarrow \gamma_1=-\delta_1$$

How the reference category is selected is irrelevant as it affects nothing but the interpretation of the coefficients associated to the dummy variables. Nevertheless, a track must be kept of which category is the reference category. Generally, a reference category is selected for convenience. One could even omit the intercept and include a dummy variable for every category. The equation then would become:

$$wage = \mu_1 male + \nu_1 female + \beta_2 educ + u \tag{8.5}$$

In the above, for women, the intercept is ν_1 and for men it is μ_1.

In the usual way, hypothesis testing is carried out. In equation 8.2, the null hypothesis of no difference what so ever between women and men is $H_0 : \delta_1 = 0$. The alternative hypothesis which claims prejudice against women is $H_1 : \delta_1 < 0$. Thus, there is need to use one sided (left) *t* test in this situation.

In applied work, a common specification has the dependent variable as the logarithm transformation ln(*y*) in models of this type. To take an example:

$$\ln(wage) = \beta_1 + \delta_1 female + \beta_2 educ + u \tag{8.6}$$

Interpretation of coefficients of dummy variables

We will now look at the dummy variable's coefficient's interpretation in a log model.

$$\ln(wage_F) = \beta_1 + \delta_1 + \beta_2 educ \tag{8.7}$$

$$\ln(wage_M) = \beta_1 + \beta_2 educ \tag{8.8}$$

In the light of the exact same education, when the equation 8.7 is subtracted at equation 8.8, the result is:

$$\ln(wage_F) - \ln(wage_M) = \delta_1 \tag{8.9}$$

If we take antilogs in 8.9 and from both sides of 8.9 we subtract 1, it gives:

$$\frac{wage_F}{wage_M} - 1 = e^{\delta_1} - 1 \tag{8.10}$$

NOTES

In other words:

$$\frac{wage_F - wage_M}{wage_M} = e^{\delta_1} - 1 \qquad (8.11)$$

Based on equation 8.11, for the same education, the proportional change between male wage and female wage will be $e^{\delta_1} - 1$. So, the precise percentage change in the hourly wages between men and women is $100 \times (e^{\delta_1} - 1)$. One can even use $100 \times \delta_1$, but not when the percentage's magnitude is high.

Let us now look at attributes with more than two categories. Here we shall look at one that has three categories.

For measuring what impact the size of an organization has on wage, let us employ a dummy variable. We will assume that there are three categories of organizations based on size, and on that, we create three variables as shown below:

$$small = \begin{cases} 1 & \text{up to 49 workers} \\ 0 & \text{in other case} \end{cases}$$

$$medium = \begin{cases} 1 & \text{from 50 to 199 workers} \\ 0 & \text{in other case} \end{cases}$$

$$large = \begin{cases} 1 & \text{more than 199 workers} \\ 0 & \text{in other case} \end{cases}$$

For explaining hourly wages through introducing the size of the organization in the model, one category will need to be dropped. In the model given below, the category *small* has been dropped.

$$wage = \beta_1 + \theta_1 medium + \theta_2 large + \beta_2 educ + u \qquad (8.12)$$

The following is the interpretation of the θ_j coefficients.

θ_1 (θ_2) represents the difference in the hourly wage between medium (large) organizations and small organizations, in the light of there being the same amount of education (and the same error term u).

The following is the model with the *small* category also included:

$$wage = \beta_1 + \theta_0 small + \theta_1 medium + \theta_2 large + \beta_2 educ + u \qquad (8.13)$$

Consider that there is a sample of six observations: small organizations' observations are one and two, the medium organizations' observation are three and four and large organizations' are five and six. In such a situation, the following would be the configuration of matrix of regressors **X**:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & educ_1 \\ 1 & 1 & 0 & 0 & educ_2 \\ 1 & 0 & 1 & 0 & educ_3 \\ 1 & 0 & 1 & 0 & educ_4 \\ 1 & 0 & 0 & 1 & educ_5 \\ 1 & 0 & 0 & 1 & educ_6 \end{bmatrix}$$

NOTES

In the above matrix, column 1 equals the sum of columns 2, 3 and 4. This shows perfect multicollinearity because of the dummy variable trap. To generalize, for an attribute with *g* categories, the model must include g-1 dummy variables along with the intercept. The reference category's intercept will be the model's overall intercept. The dummy variable coefficient for a particular group will represent the estimated variance in intercepts between that category and the reference category. In case *g* dummy variables are included along with an intercept, the dummy trap will be laid. As an alternative, it is possible to have a *g* dummy variables and drop overall intercept. So, the model would:

$$wage = \theta_0 small + \theta_1 medium + \theta_2 large + \beta_2 educ + u \qquad (8.14)$$

Such a solution should not be followed because:

- In such a model configuration, it becomes more difficult to test differences with respect to a reference category.
- The solution works only if the model has just a single unique attribute.

Qualitative and Quantitative Explanatory Variables

In the preceding example, we have taken only qualitative explanatory variables. However, the most common models in economics have some explanatory quantitative variables and other qualitative variables or regional or seasonal variables. The most common example of use of such model is the consumption function or Engel Curve (function). In the Engel function analysis, the expenditure on any commodity-group, say, foodgrain (C) depends upon total income (X). However, there may be regional differentials. Then the model can be represented as

$$C_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i \qquad .....(8.15)$$

$$D_i = 1, \text{ for Rajasthan}$$

$$= 0, \text{ otherwise}$$

Then, we have $(C_i) = (\beta_0 + \beta_2) + \beta_1 X_i + u_i$ for Rajasthan

$(C_i) = \beta_0 + \beta_1 X_i + u_i$ for otherwise.

Thus, there is difference only in the intercept term. (Figure 8.2). Obviously, in Figure 8.2, β_2 is negative.

NOTES

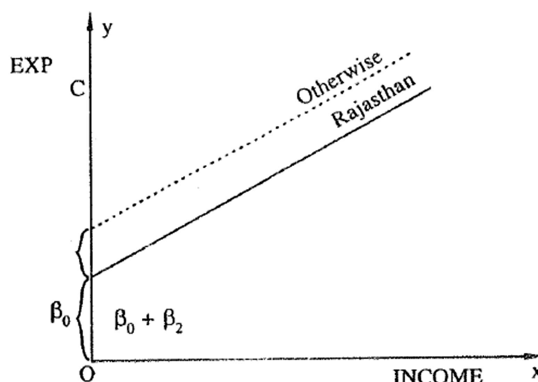


Fig. 8.2

If we estimate

$$C_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i \qquad \text{..... (8.16)}$$

then $\hat{\beta}_2$ gives the estimated regional shift (intercept = $\hat{\beta}_0 + \hat{\beta}_2$ for Rajasthan and $\hat{\beta}_0$ elsewhere). However, the slope $\hat{\beta}_1$ is common. $\hat{\beta}_1$ is OLS estimator which attempts to fit all data (Rajasthan and elsewhere). Since $\hat{\beta}_1$ is based on all data, it cannot be best fit to Rajasthan data alone nor to the other region data alone. Obviously, it is a compromise estimate.

In the dummy variable model 8.16, we estimate only three parameters. However, if we fit separate models

$$\left.\begin{array}{ll} C_i = \alpha_0 + \alpha_1 X_i + u_i & \text{Rajasthan} \\ C_i = \gamma_0 + \gamma_1 X_i + u_i & \text{Elsewhere} \end{array}\right\} \qquad \text{...(8.17)}$$

Here slopes are not constrained to be equal. Here four parameters are required to be estimated.

If we have enough data sets for both regions then we can proceed by introducing dummies in slope as well or by regarding separately.

However, if the data set for any region is very small, so that degrees of freedom are low, the separate estimate for slope coefficient will be unreliable one. Then it is better to pool the observations of the two regions (or periods) and obtain just one slope estimator.

Seasonal Factors

The dummy variable technique can be used if one has to take care of the seasonal factors. If we have quantity data on consumption (C) and national income (Y), we fit the following regression equation.

$$C_i = \beta_0 + \beta_1 Y_i + \gamma_i D_{1t} + \gamma_2 D_{2t} + \gamma_3 D_{3t} + u_i \quad ...(8.18)$$

where D_1, D_2 and D_3 are seasonal dummies defined as

$$D_{1t} = 1, \text{ for the first quarter}$$
$$= 0, \text{ otherwise}$$
$$D_{2t} = 1, \text{ for the second quarter}$$
$$= 0, \text{ otherwise}$$
$$D_{3t} = 1, \text{ for the third quarter}$$
$$= 0, \text{ otherwise}$$

The intercepts of the equation for the four quarters are

$$\text{1st quarter} = \hat{\beta}_0 + \hat{\gamma}_1$$

$$\text{2st quarter} = \hat{\beta}_0 + \hat{\gamma}_2$$

$$\text{3rd quarter} = \hat{\beta}_0 + \hat{\gamma}_3$$

$$\text{4th quarter} = \hat{\beta}_0$$

Seasonal differences are, therefore, tested by applying tests to the difference of the intercept terms. In fact, the significance of γ_1, γ_2, and γ_3 will establish the fact that there are seasonal shifts in consumption function in quarter 1, 2 and 3 as compared to quarter 4.

In decomposing into trend and seasonal factors, the following model may be used:

$$Y_t = \beta_0 + \beta_1 T + \beta_2 D_2 + \beta_3 D_3 + \beta_4 + u_i$$

where T is time, and D_2, D_3, D_4 are dummies for 2nd, 3rd and 4th quarters or seasons. Here $\hat{\beta}_0 + \hat{\beta}_1 T$ gives the trend and $\hat{\beta}_2, \hat{\beta}_3$ and $\hat{\beta}_4$ give the seasonal shifts as compared to the level in the first season (quarter).

Dummy Variables: Slopes Changing

Dummy variables can also be used for testing the differences in the slopes. Let the model be

$$C_t = \beta_0 + \beta_1 Y_t + \gamma_1 Y_t D_t + u_t \quad ....(8.19)$$

where $D_t = 1$, in war time

$= 0$, otherwise

NOTES

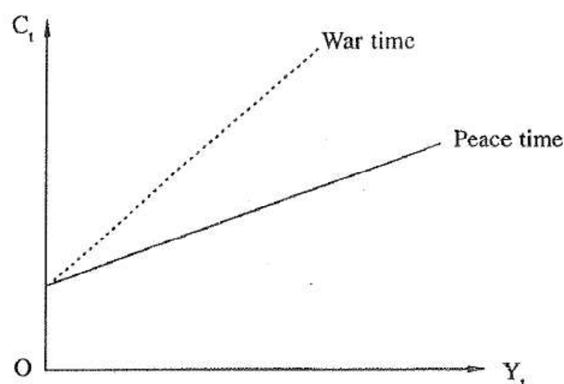


Fig. 8.3

Hence, $C_t = \beta_0 + (\beta_1\gamma_t)Y_t + u_t$ war time

$C_t = \beta_0 + \beta_1 Y_t + u_t$ peace time, (Figure 8.3). In this case, only slope is different, intercept is unchanged.

A third possibility is that both the intercept and slope may differ in war and peace time. Then our model will be:

$$C_t = \beta_0 + \beta_1 Y_t + \gamma D_1 + \gamma_1 D_t Y_t \quad ...(8.20)$$

Then, $C_t = (\beta_0 + \gamma) + (\beta_1 + \gamma_1) + Y_t + u_t$ in war time

$C_t = \beta_0 + \beta_1 Y_t + u_t$ in peace time

The intercept has changed from $\hat{\beta}_0$ to $\hat{\beta}_0 + \hat{\gamma}$ and slope from $\hat{\beta}_1$ to $\hat{\beta}_1 + \hat{\gamma}$. (Figure 8.4)

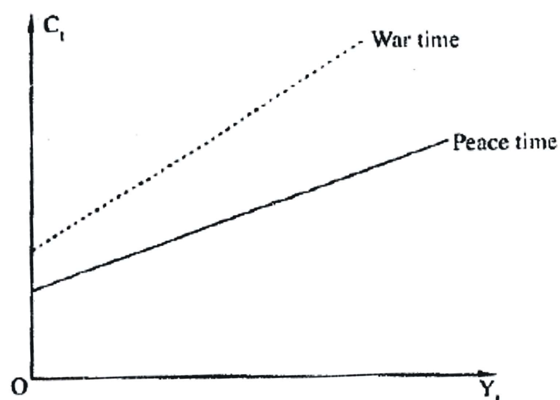


Fig. 8.4

Testing Hypothesis of Asymmetric Response

In economic theory, specially in macro economic theory, the problem of inflexibility of money wages arises. However, it is generally assumed that the money wage rate is inflexible downwards. Let us have the model.

$$N_t = \beta_0 - \beta_1 W_t + u_t$$

where N_t is employment and W_t is the money wage rate. This model implies that if money wages rises by 1 then the expected employment $E(N_t)$ will change by - β_1 *i.e.,* decrease by β_1, and if money wage rate is reduced by 1, employment will increase by β_1. The response is symmetric to change in wage rate. However, asymmetry can be introduced by inserting a dummy variable.

$$N_t = \beta_0 - \beta_1 W_t + \gamma_1 W_t D_t + u_t$$

where $D_t = 1$, when $W_i \geq W_{t-1}$

$= 0$, when $W_t < W_{t-1}$.

The test $H_0 : \gamma = 0$ is equivalent to testing for asymmetry.

Then $EN_t = \hat{\beta}_0 - (\hat{\beta}_1 - \hat{\gamma})W_t$ where $W_t \geq W_{i=1}$

$EN_t = \hat{\beta}_0 - \hat{\beta}_1 W_t$ where $W_t \leq W_{t-1}$

Binary Dependent Variable

Suppose, our dependent variable is whether or not a family has a TV set. There are several variables which influence this decision like income, wealth, occupation, age, etc. Let us take only income.

Then the model is

$$Y_t = \beta_0 + \beta_1 X_t + u_t$$

Where $Y_t = 1$, if the family has TV set

$= 0$, otherwise.

Note: Because of the special nature of dependent variables, there are some complex problems of specification of the disturbance term and estimation of parameters and interpretation of results which we do not propose to discuss here.

Test of Linearity

Dummy variable technique can also be used to test for linearity of relationships with respect to variables.

8.3 THE USE OF DUMMY VARIABLES IN SEASONAL ANALYSIS AND IN COMBINING TIME SERIES

Let us begin by understanding the time series analysis method.

Times Series Analysis Method

The time series analysis method is quite accurate where the future is expected to be similar to past. The underlying assumption in time series is that the same factors will continue to influence the future patterns of economic activity in a similar manner as in

the past. These techniques are fairly sophisticated and require experts to use these methods.

The classical approach to analysing a time series is in terms of four distinct types of variations or separate components that influence a time series. These components are:

Secular trend (or simply trend), T

The trend is a general long-term movement in the time series value of the variable (Y) over a fairly long period of time. The variable (Y) is the factor that we are interested in evaluating for the future. It could be sales, population, crime rate and so on.

Trend is a common word, popularly used in day-to-day conversation such as population trends, inflation trends, birth rate and so on. These variables are observed over a long period of time and any changes related to time are noted and calculated and a trend of these changes is established. There are many types of trends; the series may be increasing slow or increasing fast or these may be decreasing at various rates. Some remain relatively constant and some reverse their trend from growth to decline or from decline to growth over a period of time. These changes occur as a result of general tendency of the data to increase or decrease as a result of some identifiable influences.

If a trend can be determined and the rate of change can be ascertained, then tentative estimates on the same series values into the future can be made. However, such forecasts are based upon the assumption that the conditions affecting the steady growth or decline are reasonably expected to remain unchanged in the future. A change in these conditions would affect the forecasts. As an example, a time-series involving increase in population over time is illustrated in Figure 8.5.

Cyclical fluctuations (C)

The cyclical fluctuations refer to regular swings or patterns that repeat over a long period of time. The movements are considered cyclical only if they occur after time intervals of more than one year. These are the changes that take place as a result of economic booms or depressions. These may be up or down, and are recurrent in nature and have a duration of several years— usually lasting for two to ten years. These movements also differ in intensity or amplitude and each phase of movement changes gradually into the phase that follows it. Some economists believe that the business cycle completes four phases every 12 to 15 years— these four phases being: prosperity, recession, depression and recovery. However, there is no agreement on the nature or causes of these cycles.

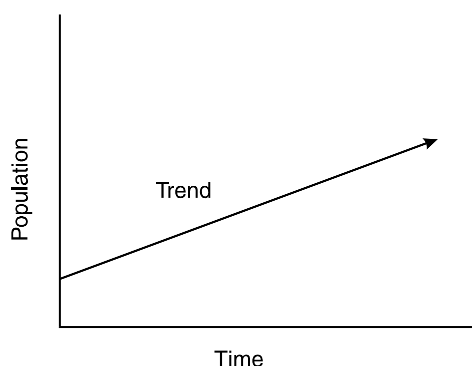


Fig. 8.5

Even though measurement and prediction of cyclical variation is very important for strategic planning, the reliability of such measurements is highly questionable due to the following reasons:

(i) These cycles do not occur at regular intervals. In the twenty-five years from 1956 to 1981 in America, it is estimated that the peaks in the cyclical activity of the overall economy occurred in August 1957, April 1960, December 1969, November 1973 and January 1980. This shows that they differ widely in timing, intensity and pattern, thus making reliable evaluation of trends very difficult.

(ii) The cyclic variations are affected by many erratic, irregular and random forces which cannot be isolated and identified separately, nor can their impact be measured accurately.

The cyclic variation for, say, revenues in an industry against time is shown graphically as follows:

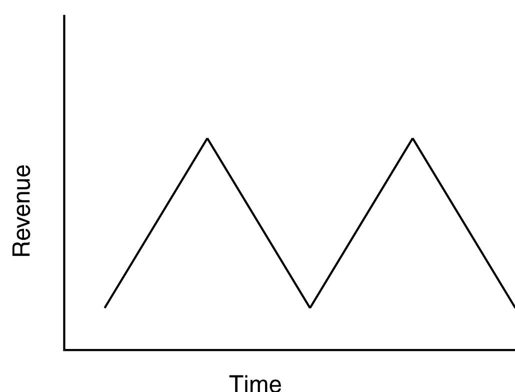


Fig. 8.6

NOTES

Seasonal variation (S)

Seasonal variation involves patterns of change that repeat over a period of one year or less. Then they repeat from year to year and they are brought about by fixed events. For example, sales of consumer items increase prior to Christmas due to gift giving tradition. The sale of automobiles in America are much higher during the last 3–4 months of the year due to the introduction of new models. This data may be measured monthly or quarterly.

Since these variations repeat during a period of 12 months, they can be predicted fairly and accurately. Some factors that cause seasonal variations are:

(i) **Season and climate:** Changes in the climate and weather conditions have a profound effect on sales. For example, the sale of umbrellas in India is always more during monsoon rainy season. Similarly, during winter, there is a greater demand for woollen clothes and hot drinks, while during summer months there is an increase in sales of fans and air conditioners.

(ii) **Customs and festivals:** Customs and traditions affect the pattern of seasonal spending. For example, Mother's Day or Valentine's Day in America see increase in gift sales preceding these days. In India, festivals such as Baisakhi and Diwali mean a big demand for sweets and candy. It is customary all over the world to give presents to children when they graduate from high school or college. Accordingly, the month of June, when most students graduate, is a time for the increase of sale for presents befitting the young.

An accurate assessment of seasonal behaviour is an aid in business planning and scheduling such as in the area of production, inventory control, personnel, advertising and so on. The seasonal fluctuations over four repeating quarters in a given year for sale of a given item is illustrated below:

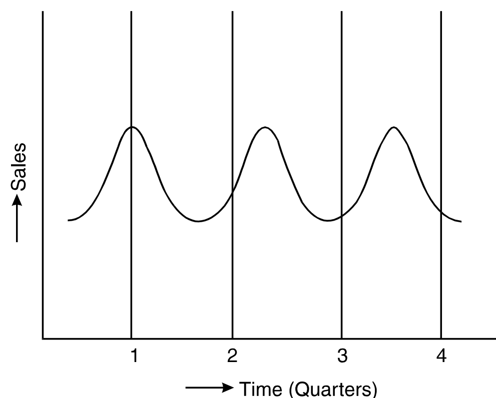


Fig. 8.7

Irregular (random) variation (I)

These variations are accidental, random or simply due to chance factors. Thus, they are wholly unpredictable. These fluctuations may be caused by such isolated incidents

NOTES

as floods, famines, strikes or wars. Sudden changes in demand or a breakthrough in a technological development may be included in this category. Accordingly, it is almost impossible to isolate and measure the value and the impact of these erratic movements on forecasting models or techniques. This phenomenon may be graphically shown as follows:

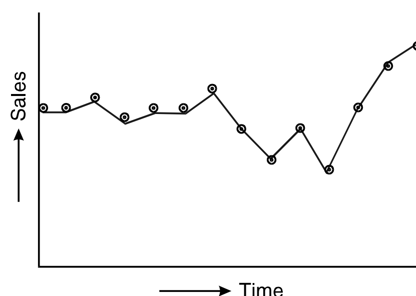


Fig. 8.8

It is traditionally acknowledged that the value of the time series (Y) is a function of the impact of variable trend (T), seasonal variation (S), cyclical variation (C) and irregular fluctuation (I). These relationships may vary depending upon assumptions and purposes. The effects of these four components might be additive, multiplicative, or combination thereof in a number of ways. However, the traditional time series analysis model is characterized by multiplicative relationship, so that:

$$Y = T \times S \times C \times I$$

The above model is appropriate for those situations where percentage changes best represent movement in the series and the components are not viewed as absolute values but as relative values.

Another approach to define the relationship may be additive, so that:

$$Y = T + S + C + I$$

This model is useful when the variations in the time series are in absolute values and can be separated and traced to each of these four parts and each part can be measured independently.

Measuring the Cyclical Effect

Cyclic variation, as we have discussed before, is a pattern that repeats over time periods longer than one year. These variations are generally unpredictable in relation to the time of occurrence, duration as well as amplitude. However, these variations have to be separated and identified. The measure we use to identify cyclical variation is the percentage of trend and the procedure used is known as the residual trend.

As we have discussed before, there are four components of time series. These are: secular trend (T), seasonal variation (S), cyclical variation (C) and irregular (or chance) variation (I). Since the time period considered for seasonal variation is less than one year, it can be excluded from the study because, when we look at

NOTES

time series consisting of annual data spread over many years, then only the secular trend, cyclical variation and irregular variation are considered.

Since secular trend component can be described by the trend line (usually calculated by line of regression), we can isolate cyclical and irregular components from the trend. Furthermore, since irregular variation occurs by chance and cannot be predicted or identified accurately, it can be reasonably assumed that most of the variations in time series left unexplained by the trend component can be explained by the cyclical component. In that respect, cyclical variation can be considered as the residual, once other causes of variation have been identified.

The measure of cyclic variation as percentage of trend is calculated as follows:

(1) Determine the trend line (usually by regression analysis).

(2) Compute the trend value Y_t for each time period (t) under consideration.

(3) Calculate the ratio Y/Y_t for each time period.

(4) Multiply this ratio by 100 to get the percentage of trend, so that:

$$\text{Percentage of trend} = \left(\frac{Y}{Y_t}\right)100.$$

Example 8.1: The following is hypothetical data for energy consumption (measured in quadrillions of BTU) in the United States from 2011 to 2016 as reported in the Statistical Abstracts of the United States.

Year	*Time Period (t)*	*Annual Energy Consumption (Y)*
2011	1	74.0
2012	2	70.8
2013	3	70.5
2014	4	74.1
2015	5	74.0
2016	6	73.9

Assuming a linear trend, calculate the percentage of trend for each year (cyclical variation).

Solution:

First we find the secular trend by the regression line method, which is given by:

$$Y_t = b_0 + b_1 t$$

where, $$b_1 = \frac{n\sum(ty) - (\sum t)(\sum y)}{n(\sum t^2) - (\sum t)^2}$$

and, $$b_0 = \bar{y} - b_1\bar{t}$$

Let us make a table for these values.

t	Y	tY	t^2
1	74.0	74.0	1
2	70.8	141.6	4
3	70.5	211.5	9
4	74.1	296.4	16
5	74.0	370.0	25
6	73.9	443.4	36
$\sum t = 21$	$\sum y = 437.3$	$\sum ty = 1536.9$	$\sum t^2 = 91$

NOTES

Substituting these values we get,

$$b_1 = \frac{6(1536.9) - (21)(437.3)}{6(91) - (21)^2}$$

$$= \frac{9221.4 - 9183.3}{546 - 441}$$

$$= \frac{38.1}{105} = 0.363$$

and, $b_0 = \bar{y} - b_1 \bar{t}$

where, $\bar{y} = \frac{\sum y}{n} = \frac{437.3}{6} = 72.88$

$$\bar{t} = \frac{21}{6} = 3.5$$

Hence, $b_0 = 72.88 - .363(3.5)$

$$= 72.88 - 1.27$$

$$= 71.61$$

Then, $Y_t = 71.61 + .363t$

Calculating the value of Y_t for each time period, we get the following table for percentage of trend $(Y/Y_t)100$.

Time Period *(t)*	*Energy Consumption* *(Y)*	*Trend* *(Y)*	*Percentage of Trend* *(Y/Y)100*
1	74.0	71.97	102.82
2	70.8	72.34	97.87
3	70.5	72.70	96.97
4	74.1	73.06	101.42
5	74.0	73.43	100.77
6	73.9	73.79	100.15

NOTES

The following graph shows the actual energy consumption (Y), trend line (Yt) and the cyclical fluctuations above and below the trend line over the time period (t) for 6 years.

Frequently, we draw a graph of cyclic variation as the percentage of trend. This process eliminates the trend line and isolates the cyclical component of the time series.

It must be understood that cyclical fluctuations are not accurately predictable, and hence we cannot predict the future cyclic variations based upon such past cyclic variations.

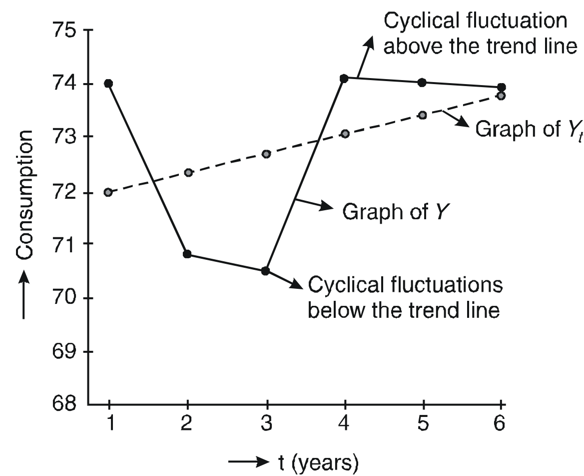


Fig. 8.9

The percentage of trend figures shows that in 2011, the actual consumption of energy was 102.82 per cent of expected consumption that year and in 2013, the actual consumption was 96.97 per cent of the expected consumption.

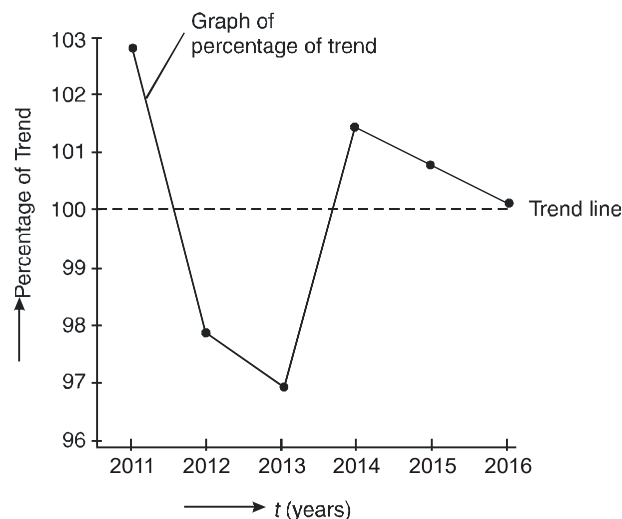


Fig. 8.10

Seasonal Variation

Seasonal variation has been defined as predictable and repetitive movement around the trend line in a period of one year or less. For the measurement of seasonal variation, the time interval involved may be in terms of days, weeks, months or quarters. Because of the predictability of seasonal trends, we can plan in advance to meet these variations. For example, study of seasonal variations in the production data makes it possible to plan for hiring of additional personnel for peak periods of production or to accumulate an inventory of raw materials or to allocate vacation time to personnel, and so on.

In order to isolate and identify seasonal variations, we first eliminate, as far as possible, the effects of trend, cyclical variations and irregular fluctuations on the time series. Some of the methods used for the measurement of seasonal variations are described as follows:

Simple Average Method

This is the simplest method of isolating seasonal fluctuations in time series. It is based on the assumption that the series contains only the seasonal and irregular fluctuations. Assume that the time series involves monthly data over a time period of, say, 5 years. Assume further that we want to find the seasonal index for the month of March. (The seasonal variation will be the same for March in every year. Seasonal index describes the degree of seasonal variation).

Then the seasonal index for the month of March will be calculated as follows:

$$\text{Seasonal Index for March } = \left(\frac{\text{Monthly average for March}}{\text{Average of Monthly Averages}}\right) \times 100$$

The following steps can be used in the calculation of seasonal index (variation) for the month of March (or any month), over the five years period, regarding the sale of cars by one distributor.

1. Calculate the average sale of cars for the month of March over the last five years.
2. Calculate the average sale of cars for each month over the five years and then calculate the average of these monthly averages.
3. Use the above formula to calculate seasonal index for March.

Let us say that the average sale of cars for the month of March over the period of five years is 360, and the average of all monthly average is 316. Then the seasonal index for March = (360/316)×100 = 113.92.

Ratio to Moving Average Method

This is the most widely used method of measuring seasonal variations. The seasonal index is based upon a mean of 100 with the degree of seasonal variation (seasonal index) measured by variations away from this base value. For example, if we look

at the seasonality of rental of row boats at the lake during the three summer months (a quarter) and we find that the seasonal index is 135 and we also know that the total boat rentals for the entire last year was 1680, then we can estimate the number of summer rentals for the row boats.

NOTES

The average number of quarterly boats rented = 1680/4 = 420.

The seasonal index, 135 for the summer quarter means that the summer rentals are 135 per cent of the average quarterly rentals.

Hence, summer rentals = 420 × (135/100) = 567.

The steps required to compute the seasonal index can be enumerated by illustrating an example.

Example 8.2: Assume that a record of rental of row boats for the last three years on a quarterly basis is given as follows:

Year	*Rentals per quarter*				*Total*
	I	II	III	IV	
1991	350	300	450	400	1500
1992	330	360	500	410	1600
1993	370	350	520	440	1680

Step 1. The first step is to calculate the four-quarter moving total for time series. This total is associated with the middle data point in the set of values for the four quarters, shown as follows.

Year	*Quarters*	*Rentals*	*Moving Total*
1991	I	350	
	II	300	
			1500
	III	450	
	IV	400	

The moving total for the given values of four quarters is 1500 which is simply the addition of the four quarter values. This value of 1500 is placed in the middle of values 300 and 450 and recorded in the next column. For the next moving total of the four quarters, we will drop the value of the first quarter, which is 350, from the total and add the value of the fifth quarter (in other words, first quarter of the next year), and this total will be placed in the middle of the next two values, which are 450 and 400, and so on. These values of the moving totals are shown in column 4 of the next table.

Step 2. The next step is to calculate the quarter moving average. This can be done by dividing the four quarter moving total, as calculated in Step 1, by 4, since there are 4 quarters. The quarters moving average is recorded in column 5 in the next table. The following entire table of calculations:

NOTES

Year	*Quarters*	*Rentals*	*Quarter Moving Total*	*Quarter Moving Average*	*Quarter Centered Moving Average*	*Percentage of Actual to Centered Moving Average*
(1)	(2)	(3)	(4)	(5)	(6)	(7)
	I	350				
	II	300				
			1500	375.0		
	III	450			372.50	120.80
			1480	370.0		
	IV	400			377.50	105.96
			1540	385.0		
1992	I	330			391.25	84.35
			1590	397.5		
	II	360			398.75	90.28
			1600	400.0		
	III	500			405.00	123.45
			1640	410.0		
	IV	410			408,75	100.30
			1630	407.5		
1993	I	370			410.00	90.24
			1650	412.5		
	II	350			416.25	84.08
			1680	420.0		
	III	520				
	IV	440				

Step 3. After the moving averages for each consecutive 4 quarters have been taken, then we centre these moving averages. As we see from the above table, the quarterly moving average falls between the quarters. This is because the number of quarters is even which is 4. If we had odd number of time periods, such as 7 days of the week, then the moving average would already be centered and the third step here would not be necessary. Accordingly, we centre our averages in order to associate each average with the corresponding quarter, rather than between the quarters. This is shown in column 6, where the centered moving average is calculated as the average of the two consecutive moving averages.

The moving average (or the centered moving average) aims to eliminate seasonal and irregular fluctuations (S and I) from the original time series, so that this average represents the cyclical and trend components of the series.

As the following graph shows for this data, the centered moving average has smoothed the peaks and troughs of the original time series.

Step 4. Column 7 in the table contains calculated entries which are percentages of the actual values to the corresponding centered moving average values. For example, the first four quarters centered moving average of 372.50 in the table

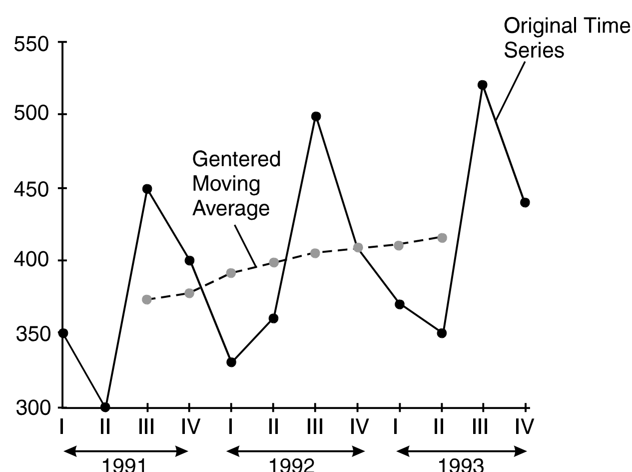
NOTES

has the corresponding actual value of 450, so that the percentage of actual value to centered moving average would be:

$$\frac{\text{Actual Value}}{\text{Centered Moving Average Value}} \times 100$$

$$= \frac{450}{372.5} \times 100$$

$$= 120.80$$



Step 5. The purpose of this step is to eliminate the remaining cyclical and irregular fluctuations still present in the values in column 7 of the table in the preceding page. This can be done by calculating the 'modified mean' for each quarter. The modified mean for each quarter of the three years time period under consideration is calculated as follows.

(a) Make a table of values in column 7 of the table in the preceding page (percentage of actual to moving average values) for each quarter of the three years as shown in the following table.

Year	*Quarter I*	*Quarter II*	*Quarter (III)*	*Quarter (IV)*
1991	–	–	120.80	105.96
1992	84.35	90.28	123.45	100.30
1993	90.24	84.08	–	–

(b) We take the average of these values for each quarter. It should be noted that if there are many years and quarters taken into consideration instead of three years as we have taken, then the highest and lowest values from each quarterly data would be discarded and the average of the remaining data would be considered. By discarding the highest and lowest values from each quarter data, we tend to reduce

the extreme cyclical and irregular fluctuations, which are further smoothed when we average the remaining values. Thus, the modified mean can be considered as an index of seasonal component. This modified mean for each quarter data is shown as follows:

$$\text{Quarter I} = \frac{84.35 + 90.24}{2} = 87.295$$

$$\text{Quarter II} = \frac{90.28 + 84.08}{2} = 87.180$$

$$\text{Quarter III} = \frac{120.80 + 123.45}{2} = 122.125$$

$$\text{Quarter IV} = \frac{105.96 + 100.30}{2} = 103.13$$

Total = 399.73

The modified means as calculated above are preliminary seasonal indices. These should average 100 per cent or a total of 400 for the 4 quarters. However, our total is 399.73. This can be corrected by the following step.

Step 6. First, we calculate an adjustment factor. This is done by dividing the desired or expected total of 400 by the actual total obtained of 399.73, so that:

$$\text{Adjustment} = \frac{400}{399.73} = 1.0007$$

By multiplying the modified mean for each quarter by the adjustment factor, we get the seasonal index for each quarter, so that:

Quarter I = 87.295 × 1.0007 = 87.356

Quarter II = 87.180 × 1.0007 = 87.241

Quarter III = 122.125 × 1.0007 = 122.201

Quarter IV = 103.13 × 1.0007 = 103.202

Total = 400.000

$$\text{Average seasonal index} = \frac{400}{4} = 100$$

(This average seasonal index is approximated to 100 because of rounding-off errors).

The logical meaning behind this method is based on the fact that the centered moving average part of this process eliminates the influence of secular trend and cyclical fluctuations ($T \times C$). This may be represented by the following expression:

$$\frac{T \times S \times C \times I}{T \times C} = S \times I$$

NOTES

where $(T \times S \times C \times I)$ is the influence of trend, seasonal variations, cyclic fluctuations and irregular or chance variations.

Thus, the ratio to moving average represents the influence of seasonal and irregular components. However, if these ratios for each quarter over a period of years are averaged, then most random or irregular fluctuations would be eliminated so that,

$$\frac{S \times I}{I} = S$$

and this would give us the value of seasonal influences.

Measuring Irregular Variation

Typically, **irregular variation** is random in nature, unpredictable and occurs over comparatively short periods of time. Because of its unpredictability, it is generally not measured or explained mathematically. Usually, subjective and logical reasoning explains such variations. For example, the Persian Gulf War, an irregular factor, resulted in increase in airline and ship travel for a number of months because of the movement of personnel and supplies. However, the irregular component can be isolated by eliminating other components from the time series data. For example, time series data contains $(T \times S \times C \times I)$ components and if we can eliminate $(T \times S \times C)$ elements from the data, then we are left with (I) component. We can follow the previous example to determine the (I) component as follows. The data presented has already been earlier provided or calculated.

Year	*Quarters*	*Rentals Time Series Values $(T \times S \times C \times I)$*	*Centered Moving Average $(T \times C)$*	*$T \times S \times C \times I/(T \times C) = S \times I$*
1991	I	350	–	–
	II	300	–	–
	III	450	372.50	1.208
	IV	400	377.50	1.060
1992	I	330	391.25	0.843
	II	360	398.75	0.903
	III	500	405.00	1.235
	IV	410	408.75	1.003
1993	I	370	410.00	0.902
	II	350	416.25	0.841
	III	520	–	–
	IV	440	–	–

NOTES

The seasonal indices for each quarter have already been calculated as:

Quarter I = 87.356

Quarter II = 87.241

Quarter III = 122.201

Quarter IV = 103.202

Then the seasonal influence is given by:

Quarter I = 87.356/100 = .874

Quarter II = 87.241/100 = .872

Quarter III = 122.201/100 = 1.222

Quarter IV = 103.202/100 = 1.032

Making another table of ($S \times I$) values and (S) values and dividing ($S \times I$) by (S) we get the values of (I) as follows:

Seasonal Adjustments

Many times we read about time series values as seasonally adjusted. This is accomplished by dividing the original time series values by their corresponding seasonal indices. These deseasonalized values allow more direct and equitable comparisons of values from different time periods. For example, in comparing the demands for rental row boats (example that we have been following), it would not be equitable to compare the demand of the second quarter (spring) with that of the third quarter (summer), when the demand is traditionally higher. However, these demand values can be compared when we remove the seasonal influence from these time series values.

The seasonally adjusted values for the demand of row boats in each quarter are based on the following values previously calculated.

Year	*Quarter*	*Rentals* ($T \times S \times C \times I$)	*(S)*	*Seasonally Adjusted Values*	*Rounded off Values*
1991	I	350	–	–	–
	II	300	–	–	–
	III	450	1.222	368.25	368
	IV	400	1.032	387.60	388
1992	I	330	0.874	377.57	378
	II	360	0.872	412.80	413
	III	500	1.222	409.16	409
	IV	410	1.032	397.29	397
1993	I	370	0.874	423.34	423
	II	350	0.872	401.38	401
	III	520	–	–	–
	IV	440	–	–	–

The seasonally adjusted value for each quarter is calculated as:

$$= \frac{\text{Original Value}}{\text{Seasonal Index}}$$

NOTES

These calculations complete the process of separating and identifying the four components of the time series, namely secular trend (T), seasonal variation (S), cyclical variation (C) and irregular variation (I).

Variation

Trend Analysis

While the chance variations are difficult to identify, separate, control or predict, a more precise measurement of trend, cyclical effects and seasonal effects can be made in order to make the forecasts more reliable. Here, we discuss the techniques that would allow us to describe trend.

When a time series shows an upward or downward long term linear trend, then regression analysis can be used to estimate this trend and project the trends into forecasting the future values of the variables involved. In regression analysis, the equation for the straight line we use to describe the linear relationship between independent variable X and dependent variable Y is:

$$Y = b_0 + b_1X$$

where, b_0 = intercept on the Y-axis and, b_1 = slope of the straight line.

In time series analysis, the independent variable is time, so we will use the symbol t in place of X and we will use the symbol Y_t in place of Y_c.

Hence, the equation for linear trend is given as:

$$Y_t = b_0 + b_1t$$

where, Y_t = forecast value of the time series in period t

b_0 = intercept of the trend line on Y-axis

b_1 = slope of the trend line

t = time period

As discussed earlier, we can calculate the values of b_0 and b_1 by the following formulae:

$$b_1 = \frac{n\sum(ty) - (\sum t)(\sum y)}{n(\sum t^2) - (\sum t)^2}, \text{ and } b_0 = \bar{y} - b_1\bar{t}$$

where, y = actual value of the time series in period time t

n = number of periods

$\bar{y}$ = average value of time series = $\frac{\sum y}{n}$

$$\bar{t} = \text{average value of } t = \frac{\sum t}{n}$$

Knowing these values, we can calculate the value of Y_t.

NOTES

Example 8.3: A car fleet owner has 5 cars which have been in the fleet for several different years. The manager wants to establish if there is a linear relationship between the age of the car and the repairs in hundreds of dollars for a given year. This way, he can predict the repair expenses for each year as the cars become older. The information for the repair costs he collected for last year on these cars is given below:

Car	*Age (t)*	*Repairs (Y)*
1	1	4
2	3	6
3	3	7
4	5	7
5	6	9

The manager wants to predict the repair expenses for next year for the two cars that are 3-years old now.

Solution:

The trend in repair costs suggests a linear relationship with the age of the car, so that the linear regression equation is given as:

$$Y_t = b_0 + b_1 t$$

where, $$b_1 = \frac{n\sum(ty) - (\sum t)(\sum y)}{n(\sum t^2) - (\sum t)^2}$$

and, $$b_0 = \bar{y} - b_1 \bar{t}$$

To calculate the various values, let us form a new table as follows:

Age of car (t)	*Repair cost (Y)*	*tY*	t^2
1	4	4	1
3	6	18	9
3	7	21	9
5	7	35	25
6	9	54	36
Totals 18	33	132	80

NOTES

Knowing that n = 5, let us substitute these values to calculate the regression coefficients b_0 and b_1.

Then, $$b_1 = \frac{5(132) - (18)(33)}{5(80) - (18)^2}$$

$$= \frac{660 - 594}{400 - 324}$$

$$= \frac{66}{76} = 0.87$$

and, $$b_0 = \bar{y} - b_1\bar{t}$$

where, $$\bar{y} = \frac{\sum y}{n} = \frac{33}{5} = 6.6$$

and, $$\bar{t} = \frac{t}{n} = \frac{18}{5} = 3.6$$

Then, $$b_0 = 6.6 - .87(3.6)$$

$$= 6.6 - 3.13$$

$$= 3.47$$

Hence $$Y_t = 3.47 + 0.87t$$

The cars that are 3-years old now will be 4-years old next year, so that t = 4.

Hence, $$Y_{(4)} = 3.47 + 0.87(4)$$

$$= 3.47 + 3.48$$

$$= 6.95$$

Accordingly, the repair costs on each car that is 3-years old now are expected to be $695.00.

Smoothing Techniques

Smoothing techniques improve the forecasts of future trends provided that the time series are fairly stable with no significant trend, cyclical or seasonal effect, and the objective is to smooth out the irregular component of the time series through the averaging process. There are two techniques that are generally employed for such smoothing.

(i) Moving averages

The concept of the moving averages is based on the idea that any large irregular component of time series at any point in time will have a less significant impact on

the trend, if the observation at that point in time is averaged with such values immediately before and after the observation under consideration. For example, if we are interested in computing the three-period moving average for any time period, then we will take the average of the value in such time period, the value in the period immediately preceding it and the value in the time period immediately following it. Let us illustrate this concept by an example.

Let the following table represent the number of cars sold in the first 6 weeks of the first two months of a year by a given dealer. Our objective is to calculate a 3 week moving average.

Week	*Sales*
1	20
2	24
3	22
4	26
5	21
6	22

The moving average for the first 3 week period is given as:

$$\text{Moving average} = \frac{29 + 24 + 22}{3} = \frac{66}{3} = 22$$

This moving average can then be used to forecast the sale of cars for week 4. Since the actual number of cars sold in week 4 is 26, we note that the error in the forecast is (26 – 22) = 4.

The calculation for the moving average for the next 3 periods is done by adding the value for week 4 and dropping the value for week 1, and taking the average for weeks 2, 3 and 4. Hence,

$$\text{Moving average} = \frac{24 + 22 + 26}{3} = \frac{72}{3} = 24$$

Then, this is considered to be the forecast of sales for week 5. Since the actual value of the sales for week 5 is 21, we have an error in our forecast of (21 – 24) = – (3).

The next moving average for weeks 3 to 5, as a forecast for week 6 is given as:

$$\text{Moving average} = \frac{22 + 26 + 21}{3} = \frac{69}{3} = 23$$

The error between the actual and the forecast value for week 6 is (22 – 23) = – (1). (Since the actual value of the sales for week 7 is not given, there is no need to forecast such values).

Our objective is to predict the trend and forecast the value of a given variable in the future as accurately as possible so that the forecast is reasonably free from random variations. To do that, we must have the sum of individual errors, as discussed above, as little as possible. However, since errors are irregular and random, it is expected that some errors would be positive in value and others negative, so that the sum of these errors would be highly distorted and would be closer to zero. This difficulty can be avoided by squaring each of the individual forecast errors and then taking the average. Naturally, the minimum values of these errors would also result in the minimum value of the 'average of the sum of squared errors'. This is shown as follows:

Week	*Time Series Value*	*Moving Average*	*Error*	*Error Squared*
1	20			
2	24			
3	22			
4	26	22	4	16
5	21	24	– 3	9
6	22	23	– 1	1

Then the average of the sum of squarred errors also known as mean squarred error and denoted by MSE is given as :

$$\text{MSE} = \frac{16+9+1}{3} = \frac{26}{3} = 8.63$$

The value of the MSE is an often-used measure of the accuracy of the forecasting method, and the method which results in the least value of the MSE is considered more accurate than others. The value of the MSE can be manipulated by varying the number of data values to be included in the moving average. For example, if we had calculated the value of the MSE by taking 4 periods into consideration for calculating the moving average, rather than 3, then the value of the MSE would be less. Accordingly, by using trial and error method, the number of data values selected for use in forecasting would be such that the resulting MSE value would be minimum.

(ii) Exponential Smoothing

In the moving average method, each observation receives the same weight. In other words, each value contributes equally towards the calculation of the moving average, irrespective of the number of time periods taken into consideration. In most actual situations, this is not a realistic assumption. Due to the dynamics of the environment over a period of time, it is more likely that the forecast for the next period would be closer to the most recent previous period than the more distant

previous period, so that the more recent value should get more weight than the previous value and so on. The exponential smoothing technique uses the moving average with appropriate weights assigned to the values taken into consideration in order to arrive at a more accurate or smoothed forecast. It takes into consideration the decreasing impact of the past time periods as we move further into the past time periods. This decreasing impact as we move down into the time period is exponentially distributed and hence the name exponential smoothing.

In this method, the smoothed value for period *t,* which is the weighted average of that period's actual value and the smoothed average from the previous period $(t-1)$, becomes the forecast for the next period $(t+1)$. Then the exponential smoothing model for time period $(t+1)$ can be expressed as follows:

$$F_{(t+1)} = \alpha Y_t + (1-\alpha)F_t$$

where, $F_{(t+1)}$ = the forecast of the time series for period $(t+1)$

Y_t = actual value of the time series in period t

α = smoothing factor $(0 \leq \alpha \leq 1)$

F_t = forecast of the time series for period t

The value of α is selected by the decision maker on the basis of degree of smoothing required. A small value of α means a greater degree of smoothing. A large value of α means very little smoothing. When $\alpha = 1,$ then there is no smoothing at all so that the forecast for the next time period is exactly the same as the actual value of times series in the current period. This can be seen by:

$$F_{(t+1)} = \alpha Y_t + (1-\alpha)F_t$$

when $\alpha = 1,$

$$F_{(t+1)} = Y_t + 0F_t = Y_t$$

The exponential smoothing approach is simple to use and once the value of α is selected, it requires only two pieces of information namely Y_t and F_t to calculate $F_{(t+1)}$.

To begin with the exponential smoothing process, we let F_t equal the actual value of the time series in period t, which is Y_1. Hence, the forecast for period 2 is written as:

$$F_2 = \alpha Y_1 + (1-\alpha)F_1$$

But since we have put $F_1 = Y_1$ hence,

$$F_2 = \alpha Y_1 + (1-\alpha)Y_1$$

$$= Y_1$$

NOTES

Let us now apply exponential smoothing method to the problem of forecasting car sales as discussed in the case of moving averages. The data once again is given as follows:

Week	*Time Series Value (Y_t)*
1	20
2	24
3	22
4	26
5	21
6	22

Let $\alpha = 0.4$

Since F_2 is calculated above as equal to $Y_1 = 20$, we can calculate the value of F_3 as follows:

$$F_3 = 0.4Y_2 + (1 - 0.4)F_2$$

Since $F_2 = Y_1$, we get

$$F_3 = .4(24) + .6(20) = 9.6 + 12$$
$$= 21.6$$

Similar values can be calculated for subsequent periods, so that:

$$F_4 = .4Y_3 + .6F_3$$
$$= .4(22) + .6(21.6)$$
$$= 8.8 + 12.96$$
$$= 21.76$$

$$F_5 = .4Y_4 + .6F_4$$
$$= .4(26) + .6(21.76)$$
$$= 10.4 + 13.056$$
$$= 23.456$$

$$F_6 = .4Y_5 + .6F_5$$
$$= .4(21) + .6(23.456)$$
$$= 8.4 + 14.07$$
$$= 22.47$$

and, $F_7 = .4Y_6 + .6F_6$

$= .4(22) + .6(22.47)$

$= 8.8 + 13.48$

$= 22.28$

Now we can compare the exponential smoothing forecast value with the actual values for the six time periods and calculate the forecast error.

Week	*Time Series Value* (Y_t)	*Exponential Smoothing Forecast Value* (F_t)	*Error* $(Y_t - F_t)$
1	20	–	–
2	24	20.000	4.0
3	22	21.600	0.4
4	26	21.760	4.24
5	21	23.456	– 2.456
6	22	22.470	– 0.47

(The value of F_7 is not considered because the value of Y_7 is not given).

Let us now calculate the value of MSE for this method with selected value of $F_{(t+1)}$ From the previous table:

Forecast errors $(Y_t - F_t)$	Squared Forecast Error $(Y_t - F_t)$
4	16
.4	.16
4.24	17.98
– 2.456	6.03
– 0.47	.22
	Total = 40.39

Then,

$$\text{MSE} = 40.39/5 = 8.08$$

The previous value of MSE was 8.67. Hence the current approach is a better one. The choice of the value α is very significant. Let us look at the exponential smoothing model again.

$$F_{(t+1)} = \alpha Y_t + (1-\alpha)F_t$$

$$= \alpha Y_t + F_t - \alpha F$$

$$= F_t + \alpha(Y_t - F_t)$$

where $(Y_t - F_t)$ is the forecast error in time period t.

The accuracy of the forecast can be improved by carefully selecting the value of α. If the time series contains substantial random variability, then a small value of α (known as smoothing factor or smoothing constant) is preferable. On the other hand, a larger value of α would be desirable for time series with relatively little random variability ($Y_t - F_t$).

Curve Fitting Methods

Curve fitting is the process of constructing a curve, or mathematical function, that has the best fit to a series of data points, possibly subject to constraints. Curve fitting can involve either interpolation, where an exact fit to the data is required, or smoothing, in which a 'smooth' function is constructed that approximately fits the data. A related topic is regression analysis, which focuses more on questions of statistical inference such as how much uncertainty is present in a curve that is fit to data observed with random errors. Fitted curves can be used as an aid for data visualization, to infer values of a function where no data are available, and to summarize the relationships among two or more variables. Extrapolation refers to the use of a fitted curve beyond the range of the observed data, and is subject to a degree of uncertainty since it may reflect the method used to construct the curve as much as it reflects the observed data.

8.4 CROSS SECTIONAL DATA

In econometrics, the cross-sectional data or a cross section of a study population is a type of data, collected by observing many subjects (such as individuals, firms, countries, or regions) at the one point or period of time. The analysis might also have no regard to differences in time. Analysis of cross-sectional data usually consists of comparing the differences among selected subjects.

Cross-sectional data differs from time series data, in which the same small-scale with factors of production aggregate entity is observed at various points in time. Another type of data, panel data (or longitudinal data), combines both cross-sectional and time series data ideas and looks at how the subjects (firms, individuals, etc.) change over a time series. Panel data differs from pooled cross-sectional data across time, because it deals with the observations on the same subjects in different times whereas the latter observes different subjects in different time periods. Panel analysis uses panel data to examine changes in variables over time and its differences in variables between selected subjects.

For example, if we want to measure current obesity levels in a population, we could draw a sample of 1,000 people randomly from that population (also known as a cross section of that population), measure their weight and height, and calculate what percentage of that sample is categorized as obese. This cross-sectional sample provides us with a snapshot of that population, at that one point

in time. Note that we do not know based on one cross-sectional sample if obesity is increasing or decreasing; we can only describe the current proportion.

NOTES

In a rolling cross-section, both the presence of an individual in the sample and the time at which the individual is included in the sample are determined randomly. For example, a political poll may decide to interview 1000 individuals. It first selects these individuals randomly from the entire population. It then assigns a random date to each individual. This is the random date that the individual will be interviewed, and thus included in the survey.

Cross-sectional data can be used in cross-sectional regression, which is regression analysis of cross-sectional data. For example, the consumption expenditures of various individuals in a fixed month could be regressed on their incomes, accumulated wealth levels, and their various demographic features to find out how differences in those features lead to differences in consumers' behaviour.

In econometrics, cross-sectional studies typically involve the use of cross-sectional regression, in order to sort out the existence and magnitude of causal effects of one independent variable upon a dependent variable of interest at a given point in time. They differ from time series analysis, in which the behaviour of one or more economic aggregates is traced through time. Cross-sectional analysis has the advantage of avoiding various complicating aspects of the use of data drawn from various points in time, such as serial correlation of residuals. It also has the advantage that the data analysis itself does not need an assumption that the nature of the relationships between variables is stable over time, though this comes at the cost of requiring caution if the results for one time period are to be assumed valid at some different point in time.

An example of cross-sectional analysis in economics is the regression of money demand—the amounts that various people hold in highly liquid financial assets—at a particular time upon their income, total financial wealth, and various demographic factors. Each data point is for a particular individual or family, and the regression is conducted on a statistical sample drawn at one point in time from the entire population of individuals or families. In contrast, an intertemporal analysis of money demand would use data on an entire country's holdings of money at each of various points in time, and would regress that on contemporaneous (or near-contemporaneous) income, total financial wealth, and some measure of interest rates.

The cross-sectional study has the advantage that it can investigate the effects of various demographic factors (age, for example) on individual differences; but it has the disadvantage that it cannot find the effect of interest rates on money demand, because in the cross-sectional study at a particular point in time all observed units are faced with the same current level of interest rates.

Check Your Progress

1. Define the nature of dummy variables.
2. Explain the qualitative explanatory variables.
3. Illustrate the testing hypothesis of asymmetric response.
4. State the binary dependent variable.
5. Elaborate on the test of linearity.
6. Interpret the time series analysis method.
7. Define the secular trend (or simply trend) *T*.
8. What do you understand by the cyclical fluctuations (C)?
9. State the irregular (random) variation (I).
10. Explain the simple average method.
11. Elaborate on the curve fitting methods.
12. What do you mean by the cross sectional data?

8.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The variables that we have used in explaining the endogenous variable have a quantitative nature. However, there are other variables of a qualitative nature that can be important when explaining the behaviour of the endogenous variable, such as sex, race, religion, nationality, and geographical region, etc.
2. The most common example of use of such model is the consumption function or Engel Curve (function). In the Engel function analysis, the expenditure on any commodity-group, say, foodgrain (C) depends upon total income (X). However, there may be regional differentials.
3. In economic theory, especially in macro-economic theory, the problem of inflexibility of money wages arises. However, it is generally assumed that the money wage rate is inflexible downwards.
4. Suppose, our dependent variable is whether or not a family has a TV set. There are several variables which influence this decision like income, wealth, occupation, and age, etc. Let us take only income. Then the model is $Y_t = \beta_0 + \beta_1 X_t + u_t$,

   where $Y_t = 1$, if the family has TV set ,

   $= 0$, otherwise.

5. Dummy variable technique can also be used to test for linearity of relationships with respect to variables.
6. The time series analysis method is quite accurate where the future is expected to be similar to past. The underlying assumption in time series is that the same factors will continue to influence the future patterns of economic activity in a similar manner as in the past. These techniques are fairly sophisticated and require experts to use these methods.
7. The trend is a general long-term movement in the time series value of the variable (Y) over a fairly long period of time. The variable (Y) is the factor that we are interested in evaluating for the future. It could be sales, population, crime rate and so on. Trend is a common word, popularly used in day-to-day conversation such as population trends, inflation trends, birth rate and so on.
8. The cyclical fluctuations refer to regular swings or patterns that repeat over a long period of time. The movements are considered cyclical only if they occur after time intervals of more than one year. These are the changes that take place as a result of economic booms or depressions.
9. These variations are accidental, random or simply due to chance factors. Thus, they are wholly unpredictable. These fluctuations may be caused by such isolated incidents as floods, famines, strikes or wars. Sudden changes in demand or a breakthrough in a technological development may be included in this category.
10. This is the simplest method of isolating seasonal fluctuations in time series. It is based on the assumption that the series contains only the seasonal and irregular fluctuations.
11. Curve fitting is the process of constructing a curve, or mathematical function that has the best fit to a series of data points, possibly subject to constraints. Curve fitting can involve either interpolation, where an exact fit to the data is required, or smoothing, in which a 'Smooth' function is constructed that approximately fits the data.
12. In econometrics, the cross-sectional data or a cross section of a study population is a type of data, collected by observing many subjects (such as individuals, firms, countries, or regions) at the one point or period of time. The analysis might also have no regard to differences in time. Analysis of cross-sectional data usually consists of comparing the differences among selected subjects.

8.6 SUMMARY

- The variables that we have used in explaining the endogenous variable have a quantitative nature. However, there are other variables of a qualitative

NOTES

nature that can be important when explaining the behaviour of the endogenous variable, such as sex, race, religion, nationality, and geographical region, etc.

- Oftentimes, qualitative factors are found to be binary information, for example, such as whether an individual is a female or male, single or not, and so forth. Where qualitative factors are found as dichotomous information, it becomes possible to capture the relevant information through defining a zero-one/binary variable.
- The most common example of use of such model is the consumption function or Engel Curve (function). In the Engel function analysis, the expenditure on any commodity-group, say, foodgrain (C) depends upon total income (X). However, there may be regional differentials.
- The dummy variable technique can be used if one has to take care of the seasonal factors.
- In economic theory, especially in macro-economic theory, the problem of inflexibility of money wages arises. However, it is generally assumed that the money wage rate is inflexible downwards.
- Because of the special nature of dependent variables, there are some complex problems of specification of the disturbance term and estimation of parameters and interpretation of results which we do not propose to discuss here.
- Dummy variable technique can also be used to test for linearity of relationships with respect to variables.
- The time series analysis method is quite accurate where the future is expected to be similar to past. The underlying assumption in time series is that the same factors will continue to influence the future patterns of economic activity in a similar manner as in the past. These techniques are fairly sophisticated and require experts to use these methods.
- The trend is a general long-term movement in the time series value of the variable (Y) over a fairly long period of time. The variable (Y) is the factor that we are interested in evaluating for the future. It could be sales, population, crime rate and so on.
- Trend is a common word, popularly used in day-to-day conversation such as population trends, inflation trends, birth rate and so on.
- There are many types of trends; the series may be increasing slow or increasing fast or these may be decreasing at various rates. Some remain relatively constant and some reverse their trend from growth to decline or from decline to growth over a period of time. These changes occur as a result of general tendency of the data to increase or decrease as a result of some identifiable influences.
- If a trend can be determined and the rate of change can be ascertained, then tentative estimates on the same series values into the future can be made.

- The cyclical fluctuations refer to regular swings or patterns that repeat over a long period of time. The movements are considered cyclical only if they occur after time intervals of more than one year. These are the changes that take place as a result of economic booms or depressions.
- The cyclic variations are affected by many erratic, irregular and random forces which cannot be isolated and identified separately, nor can their impact be measured accurately.
- These variations are accidental, random or simply due to chance factors. Thus, they are wholly unpredictable. These fluctuations may be caused by such isolated incidents as floods, famines, strikes or wars. Sudden changes in demand or a breakthrough in a technological development may be included in this category.
- Cyclic variation are generally unpredictable in relation to the time of occurrence, duration as well as amplitude. However, these variations have to be separated and identified.
- Curve fitting is the process of constructing a curve, or mathematical function that has the best fit to a series of data points, possibly subject to constraints. Curve fitting can involve either interpolation, where an exact fit to the data is required, or smoothing, in which a 'Smooth' function is constructed that approximately fits the data.
- In econometrics, the cross-sectional data or a cross section of a study population is a type of data, collected by observing many subjects (such as individuals, firms, countries, or regions) at the one point or period of time.
- The analysis might also have no regard to differences in time. Analysis of cross-sectional data usually consists of comparing the differences among selected subjects.

8.7 KEY WORDS

- **Dummy variables:** The variables that we have used in explaining the endogenous variable have a quantitative nature. However, there are other variables of a qualitative nature that can be important when explaining the behaviour of the endogenous variable, such as sex, race, religion, nationality, and geographical region, etc.
- **Qualitative explanatory variables:** The most common example of use of such model is the consumption function or Engel Curve (function). In the Engel function analysis, the expenditure on any commodity-group, say, foodgrain (C) depends upon total income (X).
- **Test of linearity:** Dummy variable technique can also be used to test for linearity of relationships with respect to variables.

NOTES

- **Time series analysis method:** The time series analysis method is quite accurate where the future is expected to be similar to past. The underlying assumption in time series is that the same factors will continue to influence the future patterns of economic activity in a similar manner as in the past.
- **The trend:** Trend is a common word, popularly used in day-to-day conversation such as population trends, inflation trends, birth rate and so on.
- **Cyclical fluctuations:** The cyclical fluctuations refer to regular swings or patterns that repeat over a long period of time. The movements are considered cyclical only if they occur after time intervals of more than one year.
- **Irregular variation:** These variations are accidental, random or simply due to chance factors. Thus, they are wholly unpredictable.
- **Simple average method:** This is the simplest method of isolating seasonal fluctuations in time series. It is based on the assumption that the series contains only the seasonal and irregular fluctuations.
- **Curve fitting method:** Curve fitting is the process of constructing a curve, or mathematical function that has the best fit to a series of data points, possibly subject to constraints.
- **Cross-sectional data:** In econometrics, the cross-sectional data or a cross section of a study population is a type of data, collected by observing many subjects (such as individuals, firms, countries, or regions) at the one point or period of time.

8.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Explain the nature of dummy variables.
2. Define the qualitative explanatory variables.
3. Interpret the testing hypothesis of asymmetric response.
4. Elaborate on the binary dependent variable.
5. State the test of linearity.
6. Illustrate the time series analysis method.
7. Explain the secular trend (or simply trend) T.
8. What do you mean by the cyclical fluctuations (C)?
9. Define the irregular (random) variation (I).
10. State the simple average method.

11. What do you understand by the curve fitting methods?
12. Elaborate on the cross sectional data.

Long-Answer Questions

NOTES

1. Discuss briefly the regression on dummy variables with the help of examples.
2. Analyse the nature of dummy variables. What is the use of dummy variables for capturing the effect of qualitative factors?
3. Explain the use of dummy variables in seasonal analysis and in combining time series.
4. Differentiate between the simple average method and curve fitting methods.
5. Briefly define the cross sectional data. Give appropriate examples.

8.9 FURTHER READINGS

Johnston, J. and John DiNARDO. 1997. *Econometric Methods*, Fourth Edition. New Delhi: Tata McGraw-Hill.

Koutsoyiannis, A.1977. *Theory of Econometrics,* Second Edition. London: The Macmillan Press Ltd.

Özdemir, Durmuº. 2016. *Applied Statistics for Economics and Business,* Second Edition. Izmir (Turkey): Springer.

Maddala, G. S. 1992. *Introduction to Econometrics,* Second Edition. New York: Macmillan Publishing Company.

Pindyck, R. S and D. L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts,* Fourth Edition. New York: McGraw Hill.

Goldberger, A. S. 1998. *Introductory Econometrics***.** Cambridge: Harvard University Press.

Levine, David M., Timothy C. Krehbiei, Mark L. Berenson and P. K. Viswanathan. 2009. *Business Statistics*, Fifth Edition. New Delhi: Pearson Education.

Webster, Allen L. 1998. *Applied Statistics for Business and Economics,* Third Edition. New Delhi: Tata McGraw-Hill.

UNIT 9 PROBLEMS OF INFERENCE

NOTES

Structure

9.0 INTRODUCTION

In econometrics, statistical inference makes propositions about a population, using data drawn from the population with some form of sampling. Given a hypothesis about a population, for which we wish to draw inferences, statistical inference consists of (first) selecting a statistical model of the process that generates the data and (second) deducing propositions from the model. Konishi and Kitagawa state, "The majority of the problems in statistical inference can be considered to be problems related to statistical modelling". Relatedly, Sir David Cox has said, "How the translation from subject-matter problem to statistical model is done is often the most critical part of an analysis"?

Statistical inference is the process of using data analysis to infer properties of an underlying distribution of probability. Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population.

Inferential statistics can be contrasted with descriptive statistics. Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population. In machine learning, the term inference is sometimes used instead to mean "make a prediction, by evaluating an already trained model"; in this context inferring properties of the model is referred to as training or learning (rather than inference), and using a model for prediction is referred to as inference (instead of prediction); see also predictive inference.

Any statistical inference requires some assumptions. A statistical model is a set of assumptions concerning the generation of the observed data and similar data. Descriptions of statistical models usually emphasize the role of population

quantities of interest, about which we wish to draw inference. Descriptive statistics are typically used as a preliminary step before more formal inferences are drawn.

In this unit, you will study about the problem of inference, the normality assumption, hypothesis testing about individual partial regression coefficients, and testing the overall significance of the sample regression.

9.1 OBJECTIVES

After going through this unit, you will be able to:

- Elaborate on the problem of inference
- Comprehend the normality assumption
- Analyse the hypothesis testing about individual partial regression coefficients
- Define the testing the overall significance of the sample regression

9.2 THE NORMALITY ASSUMPTION

In multiple regression, the assumption needing a normal distribution applies only to the disturbance term, not to the independent variables as is often supposed. Perhaps, the confusion about this assumption derives from difficulty understanding what this disturbance term refers to – simply put, it is the random error in the relationship between the independent variables and the dependent variable in a regression model.

In actual, each case in the sample has a different random variable which encompasses all the “Noise” that accounts for differences in the observed and predicted values produced by a regression equation. The distribution of this disturbance term or noise for all cases in the sample that should be normally distributed. In econometrics, normality tests are used to determine if a data set is well-modelled by a normal distribution and to compute how likely it is for a random variable underlying the data set to be normally distributed.

More precisely, the tests are a form of model selection, and can be interpreted several ways, depending on one’s interpretations of probability:

More precisely, the tests are a form of model selection, and can be interpreted several ways, depending on one’s interpretations of probability:

- In descriptive statistics terms, one measures a goodness of fit of a normal model to the data – if the fit is poor then the data are not well modelled in that respect by a normal distribution, without making a judgment on any underlying variable.
- In frequentist statistics statistical hypothesis testing, data are tested against the null hypothesis that it is normally distributed.

NOTES

- In Bayesian statistics, one does not "Test Normality" per se, but rather computes the likelihood that the data come from a normal distribution with given parameters μ,σ (for all μ,σ), and compares that with the likelihood that the data come from other distributions under consideration, most simply using a Bayes factor (giving the relative likelihood of seeing the data given different models), or more finely taking a prior distribution on possible models and parameters and computing a posterior distribution given the computed likelihoods.

A normality test is used to determine whether sample data has been drawn from a normally distributed population (within some tolerance). A number of statistical tests, such as the Student's t-test and the one-way and two-way ANOVA require a normally distributed sample population.

Assumptions deliver a way for economists to shorten the economic data and make them easier to analyse. An assumption allows an economist to break down a complex process into a simple manner. Hence, better simplification will allow the economists to emphasis only on the most relevant variables.

Observation of data is perilous for economists because they take the results and interpret them in a significant way. Cause and effect relationships are used to establish economic theories and principles. If a theory accepted universally, it becomes a law. In general, a law is always considered to be true. The scientific method delivers the framework necessary for the progression of economic study. All economic theories, principles, and laws are generalizations or abstractions. Through the use of the scientific method, economists are able to break down complex economic scenarios in order to gain a deeper understanding of critical data.

There are two approaches to statistical inference: model-based inference and design-based inference. Both approaches rely on some statistical model to represent the data-generating process. In the model-based approach, the model is taken to be initially unknown, and one of the goals is to select an appropriate model for inference. In the design-based approach, the model is taken to be known, and one of the goals is to ensure that the sample data are selected randomly enough for inference.

Given that the validity of any conclusion drawn from a statistical inference depends on the validity of the assumptions made, it is clearly important that those assumptions should be reviewed at some stage. Some instances—for example where data are lacking—may require that researchers judge whether an assumption is reasonable. Researchers can expand this somewhat to consider what effect a departure from the assumptions might produce. Where more extensive data are available, various types of procedures for statistical model validation are available—e.g., for regression model validation.

9.3 HYPOTHESIS TESTING ABOUT INDIVIDUAL PARTIAL REGRESSION COEFFICIENTS

NOTES

A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. Hypothesis refers to a provisional idea whose merit needs evaluation, but having no specific meaning. Though it is often referred as a convenient mathematical approach for simplifying cumbersome calculation. Setting up and testing hypothesis is an integral art of statistical inference. Hypotheses are often statements about population parameters like variance and expected value. During the course of hypothesis testing some inference about population like the mean and proportion are made. Any useful hypothesis will enable predictions by reasoning including deductive reasoning. According to Karl Popper, a hypothesis must be falsifiable and that a proposition or theory cannot be called scientific if it does not admit the possibility of being shown false. Hypothesis might predict outcome of an experiment in a lab, setting the observation of a phenomenon in nature. Thus, hypothesis is a explanation of a phenomenon proposal suggesting a possible correlation between multiple phenomena.

Hypothesis is put forward as a proposition. It may even be a set of more than one proposition. A proposition is the antecedent of a conditional proposition which may be an assumption or a guess. It is something yet to be proved, but taken to be temporarily true.

Hypothesis is applied in Natural Sciences, as a tentative theory provisionally adopted to explain few facts that provide guidance in proving other facts. This is often known as a working hypothesis.

A hypothesis may be a proposal that explains certain facts based on certain observations. It may be a message which is an opinion and it may be based on certain evidences which may be incomplete.

Formalized hypotheses have two variables, independent and dependent. The independent variable is the person, may be the scientist, who is going to put the hypothesis and the dependent variable is one that the person observes.

A good hypothesis has three characteristics. It is testable as it is based on sound rationale, and it is practical and ethical to conduct the test.

Statistical hypothesis can not ascertain the truth of the population parameter. To do this, truth table for entire population is required to be examined which is time consuming and impractical. Researchers examine a random sample and after judging its consistency, the hypothesis is accepted.

Statistical hypotheses are of two types.

- **Null Hypothesis:** It signifies that sample observations result purely from chance. Its notation is H_0.
- **Alternative Hypothesis:** It signifies that sample observations are influenced by some cause of non-random nature. Its notation is H_1 or H_a.

NOTES

Type I Error: A Type I error is defined as rejection of a null hypothesis when it is true. It signifies wrong decision. Probability of a Type I error is known as **significance level**, named alpha and denoted by Greek letter α.

Type II Error: A Type II error is the acceptance of null hypothesis when it is false. This too signifies wrong decision. Probability of a Type II error is known as **Beta**, and denoted by Greek letter β. Complement of this probability is power of the test.

The characteristics of hypothesis are:

- **Clear and Accurate:** Hypothesis should be clear and accurate so as to draw a consistent conclusion.
- **Statement of Relationship between Variables:** If a hypothesis is relational, it should state the relationship between different variables.
- **Testability:** A hypothesis should be open to testing so that other deductions can be made from it and can be confirmed or disproved by observation. The researcher should do some prior study to make the hypothesis a testable one.
- **Specific with Limited Scope:** A hypothesis, which is specific with limited scope, is easily testable than a hypothesis with limitless scope. Therefore, a researcher should pay more time to do research on such kind of hypothesis.
- **Simplicity:** A hypothesis should be stated in the most simple and clear terms to make it understandable.
- **Consistency:** A hypothesis should be reliable and consistent with established and known facts.
- **Time Limit:** A hypothesis should be capable of being tested within a reasonable time. In other words, it can be said that the excellence of a hypothesis is judged by the time taken to collect the data needed for the test.
- **Empirical Reference:** A hypothesis should explain or support all the sufficient facts needed to understand what the problem is all about.

A hypothesis is a statement or assumption concerning a population. For the purpose of decision-making, a hypothesis has to be verified and then accepted or rejected. This is done with the help of observations. We test a sample and make a decision on the basis of the result obtained. Decision-making plays significant role in different areas such as marketing, industry and management.

Statistical Decision-Making

Testing a statistical hypothesis on the basis of a sample enables us to decide whether the hypothesis should be accepted or rejected. The sample data enable us to accept or reject the hypothesis. Since the sample data give incomplete information

about the population, the result of the test need not be considered to be final or unchallengeable. The procedure, on which the basis of sample results, enables to decide whether a hypothesis is to be accepted or rejected. This is called Hypothesis Testing or Test of Significance.

Note 1: A test provides evidence, if any, against a hypothesis, usually called a null hypothesis. The test cannot prove the hypothesis to be correct. It can give some evidence against it.

The test of hypothesis is a procedure to decide whether to accept or reject a hypothesis.

Note 2: The acceptance of a hypotheses implies if there is no evidence from the sample that we should believe otherwise.

The rejection of a hypothesis leads us to conclude that it is false. This way of putting the problem is convenient because of the uncertainty inherent in the problem. In view of this we must always briefly state a hypothesis that we *hope to reject.*

A hypothesis stated in the hope of being rejected is called a *null hypothesis* and is denoted by H_0.

If H_0 is rejected, it may lead to the acceptance of an alternative hypothesis denoted by H_1.

For example, New fragrance soap is introduced in the market. The null hypothesis H_0, which may be rejected, is that the new soap is not better than the existing soap.

Similarly, a dice is suspected to be rolled. Roll the dice a number of times to test.

The Null Hypothesis H_0: $p = 1/6$ for showing six.

The Alternative hypothesis H_1: $p \neq 1/6$.

For example, Skulls found at an ancient site may all belong to race X or race Y on the basis of their diameters. We may test the hypothesis that the mean is μ of the population from which the present skulls came. We have the hypotheses.

$H_0 : \mu = \mu_x, H_1 : \mu = \mu_y$

Here, we should not insist on calling either hypothesis null and the other alternative since the reverse could also be true.

Simple and Composite Hypotheses

A simple hypothesis is one that specifies complete population distribution.

For example,

1. H_0: $X \sim Bi(150,1/2)$, i.e., p is given ($p = ½$)
2. H_0: $X \sim N(8,30)$, i.e., μ and s^2 are given.

NOTES

In composite hypothesis population distribution is *not* specified completely. The following examples will make the concept clear.

1. X ~ Bi(150,p) and H_1: p > 0.5
2. X ~ N(0, s^2) and H_1: s^2 is not specified.

Composite hypothesis has one or more free parameters. We can cite example on a hypothesis that the decay of a particle of a radioactive element is purely exponential with unknown lifetime. This is a composite hypothesis.

Testing of Simple Hypothesis

Many applications in language engineering require testing of hypotheses. Suppose we have to differentiate between a person's 'reading a speech' and 'giving spontaneous speech'. If testing aims at differentiating between read and spontaneous speech with respect to selected statistics and the criteria is put as mean vowel duration in the two conditions where speech was recorded. This is simple hypothesis testing since it involves a parameter of a single population.

Concept involved in such a testing is; making alternative assertions about the likely outcome of an analysis. One assertion is; there is no difference between the two conditions. This is null hypothesis, denoted as H_0 which asserts that the mean tone unit duration in the read speech is the same as that in the spontaneous speech.

There may be other assertions which are called *alternative hypotheses*, denoted *as* H_1 or H_a. An alternative hypothesis may assert that the tone unit duration of the read speech will be less than that of the spontaneous speech. A second may be the converse of it.

One-Tailed or Two-Tailed Hypotheses

The decision on selection of alternate hypotheses, to propose, depends on factors leading the language engineer to note differences in one direction or the other. These instances are referred to as *one-tailed or two-tailed hypotheses* depending on whether differentiation is being done for one direction or both the directions. Here, large differences between the means of the read and spontaneous speech, regardless of the direction followed may give evidence in favour of the alternative hypothesis.

It is important to make distinction between one-tailed and a two-tailed test. This affects the decision to assert a significant difference and this supports the null hypothesis. One-tailed test, needs smaller differences between means in comparison to that needed for a two-tailed test.

Time is noted for both the cases of read and spontaneous speech, and the samples are from the same speaker. But if it is required to have a related groups test instead of an independent group, the *t*-statistic is calculated as:

$$t = \frac{\text{Mean of condition}_1 - \text{Mean of condition}_2}{S.E.\text{ of differences}}$$

NOTES

A statistic collected for 20 speakers records a mean tone unit duration of 38.6 centiseconds and the spontaneous speech 33.4 centiseconds and the standard deviation of the difference between the means is 2.65, then *t* value is 1.96 [=(38.6 – 33.4)/2.65]. This *t* value is used to establish whether two sample means differing so much, might have come from the same (null hypothesis) or different (alternate hypothesis) distributions.

Decision rules are formulated to assess a level of support for the alternate hypothesis. Basically this involves stipulations that assumes the samples from the same distribution. But if the probability of the means differ so much, is so little that one may think of an alternative conclusion that the samples are drawn from different populations.

Such a stipulation is done at discrete probability levels. If there is a less than 5% chance of samples belonging to the same distribution, then the hypothesis that the samples were drawn from different distributions is supported as alternative hypothesis at that level of significance. If there is a chance, more than 5 per cent, that the samples are drawn from the same distribution, the null hypothesis is supported. In the worked example, with 19 degrees of freedom, a *t* value of 1.96 does not lead to the conclusion that samples are drawn from different populations, thus the null hypothesis is accepted.

Test Statistic

These are quantified parameters calculated from sample of data, which is used to decide the acceptance or rejection of the null hypothesis.

Test statistic of a hypothesis test is given by:

$$Z = \frac{Y - \mu_0}{\sigma_7} = \frac{Y - \mu_0}{\sigma / \sqrt{n}}$$ where symbols have their usual meaning.

Y stands for mean, μ_0 for claim level = H_0, σ = Standard deviation. σ^2 is variance and Z gives a test statistic.

Critical Value(s)

This is defined as a threshold value used to decide the criteria of accepting or rejecting null hypothesis. This depends on the significance level of the test.

Significance Level

This is given as a fixed probability of wrongfully rejecting the null hypothesis H_0 and is the probability of a type I error. This decision is taken by a person or agency carrying out the investigation.

Critical Region

This region is defined as a set of test statistic leading to rejection of the null hypothesis in a hypothesis test. For this the sample space is split into two mutually exclusive regions. This region provides basis to reject the null hypothesis H_0.

NOTES

P-Value

This is a probability value, and known as p-value. This is the probability of getting extreme value of the test statistic in comparison to that observed by chance alone, provided the null hypothesis H_0, is true. It is the probability of wrongly rejecting the null hypothesis.

Power

The power of a statistical hypothesis test measures the ability of such a test to reject the null hypothesis when it is actually false. Thus it is power to make a correct decision. This can be retold as the power of not committing a type II error. It is given by subtracting the probability of a type II error from 1. Mathematically it is given as:

Power = 1 – P(type II error) = (1–β)?

The maximum power a test can have is 1, and the minimum is 0. Ideally we want a test to have high power, close to 1. Normally 0.8 is considered as good for correct decision-making.

Power of a statistical Test

A statistical hypothesis test is a test in which a hypothesis is tested. It analyses gathered data to decide on hypothesis. In decision making such analysis is highly desired. It calculates values of some key variables and compares with a critical value for which hypothesis is assumed to be true. In case value of such variable(s) are far away from the critical value, it rejects the hypothesis.

Hypothesis Tests

Two hypotheses are made; one is 'Null hypothesis' the other is called 'alternative hypothesis'. A formal process is followed defining some standards and then decide whether to accept the hypothesis or reject it. This is known as hypothesis testing which has four steps.

1. **Statements:** Make two statements which are hypotheses. These are: 'null hypothesis' and 'alternative hypotheses'. Null hypothesis is denoted by H_0 and alternative hypothesis by H_a. H_0 and H_a are mutually exclusive. If H_0 is true then H_a must be false and vice-versa.
2. **Deciding Strategy:** Set formula to compute salient parameters and chalk out analysis plan that describes the use of sample data and compute the critical parameters to decide whether to accept or reject the null hypothesis.
3. **Sample Data Analysis:** Values of critical parameters like mean, proportion, t-score, z-score, and other parameters as decided by the researcher.
4. **Interpretation on Results:** Based on decision rules, null hypothesis is accepted or rejected.

NOTES

In any test, error can take place and this is also true for statistical tests. Statistic defines two types of errors; type I and Type II.

- **Type I Error:** A Type I error is defined as rejection of a null hypothesis when it is true. It signifies wrong decision. Probability of a Type I error is known as **significance level**, named alpha and denoted by Greek letter α.
- **Type II Error:** A Type II error is acceptance of null hypothesis when it is false. This too signifies wrong decision. Probability of a Type II error is known as **Beta**, and denoted by Greek letter β. Complement of this probability is power of the test.

The ***critical region*** is defined as a set of all outcomes of a hypothesis test that leads to the rejection of a null hypothesis and acceptance of an alternative hypothesis accepted and is denoted by C.

Power of a Test

A test is more powerful if it is able to worked out a criteria that directs to a clear decision. Probability of producing significant difference for such decisions is known as power of test. This difference must be found at a significance level decided by the researcher to asses the power of the test.

This probability signifies the power of making correct decision. It is complement of probability of occurrence of type II error, β **(Beta).** Thus, power of a test = $(1-\beta)$.

Power of a statistical test is affected by differences between the sample size and the specified significance level. A power of 1 is ideal, but in practice 0.80 or more is taken as good value to decide departure from the null hypothesis.

A significance criterion is a statement of unlikelihood of a result. Here the null hypothesis is considered significant. Commonly used criteria take probabilities as 0.05, 0.01 and 0.001. The power of a test is increased by weakening the significance level, putting this under a narrow limit. This increases the chance of obtaining a statistically significant result by rejecting the null hypothesis correctly and thus taking a correct decision. But it increases the risk of a Type I error.

There is no formal standard for power. In practice, a power of 0.80 or more is considered good to detect a reasonable departure from the null hypothesis.

Size/Significance Level of a Test (α)

Significance level in simple hypothesis test is the probability of *incorrectly* rejecting the null hypothesis. In a composite hypothesis it is the upper bound of the probability that serves the basis of rejecting the null hypothesis.

The greatest power for a given *significance level* is known as **most powerful test.**

NOTES

A test that has greatest *power* for all values of the parameter under test is **Uniformly Most Powerful (UMP) to test**.

When power of test is near unity, the test is consistent. This is termed 'consistent test'.

Unbiased test for a H_a in which the probability of rejection of H_0 > significance level for H_a to be true; and d^0 the significance level when H_0 is true.

Uniformly Most Powerful Unbiased (UMPU) test, is a UMP in the set of all unbiased tests.

Testing Hypothesis

A claim or hypothesis about the values or population parameters is known as the Null Hypothesis and is written as H_0. In the case of the above discussed situation, our assumption that a butler is innocent would form the null hypothesis and would be stated as follows:

$$H_0 = \text{The butler is innocent}$$

This hypothesis is then tested with the available evidence and the decision is made whether to accept this hypothesis or reject it. If this hypothesis is rejected, then we accept the alternate hypothesis which is that the butler is not innocent. This alternate hypothesis is denoted as H_1 and is stated as:

$$H_1 = \text{The butler is not innocent}$$

The process involves testing of the null hypothesis. If the null hypothesis is rejected, then the alternate hypothesis is accepted. It should be noted that the acceptance of the alternate hypothesis does not mean that it is correct. It simply means that there is not enough evidence to be reasonably sure that the null hypothesis is acceptable.

As already explained, there are two types of errors that can be used in making decisions regarding accepting or rejecting the null hypothesis. The first type of error, known as Type I error is used when the null hypothesis is rejected even if it is true. The second type of error, known as Type II error is used when a null hypothesis is accepted even if it was not true and should have been rejected.

In statistical hypothesis testing and decision-making about the values of population parameters as defined by the sample statistics, the null hypothesis asserts that there is no true difference between the sample statistics and the corresponding population parameter under consideration and if indeed there is any visible difference, it is considered to be due to natural fluctuations in sampling.

To conclude we say that,

- *Null Hypothesis* H_0– An assertion about the population parameter that is being tested by the sample results.
- *Alternate Hypothesis* H_1 – A claim about the population parameter that is accepted when the null hypothesis is rejected.

- *Type I Error*—An error made in rejecting the null hypothesis, when in fact it is true.
- *Type II Error*–An error made in accepting the null hypothesis, when in fact it is false.

Type I error is denoted by α (Alpha) and is expressed as a probability of rejecting a true hypothesis. It is also known as the level of significance. $1 - \alpha$ expresses the level of confidence. For example, $\alpha = 0.05$ means that the confidence level is 95% or 0.95.

Type II error is denoted by β (Beta) and is expressed as the probability of accepting a false hypothesis. It is desirable to have the β value as low as possible for its value reflects the power of the test being performed and a low β value indicates that the test of significance is powerful and reliable.

Procedure For Hypothesis Testing

The general procedure for hypothesis testing consists of the following steps:

1. ***State the null hypothesis as well as the alternate hypothesis.*** This means stating the assumed value of the population parameter which is to be tested. For example, suppose that we want to test the hypothesis that the average IQ of our college students is 130. Then this would become our null hypothesis and the alternate hypothesis would be that this average IQ is not 130. These statements are expressed as follows:

$$H_0 : \mu = 130$$
$$H_1 : \mu \neq 130$$

2. ***Establish a level of significance prior to sampling.*** The level of significance signifies the probability of committing Type I error α and is generally taken as equal to 0.05, which really means that after the hypothesis has been tested and a decision is made, we will still be making an error in rejecting the null hypothesis when in fact it is true, 5% of the time. Sometimes the value α is established as 0.01, but it is at the discretion of the investigator to select its value, depending upon the sensitivity of the study.
3. ***Determine a suitable test statistic.*** This means the choice of appropriate probability distribution to use with the particular available information under consideration. The normal distribution using the Z score table or the *t*-distribution is most often used.
4. ***Define the rejection (critical) regions.*** The critical region will be established on the basis of the choice of the value of the level of significance α. For example, if we select the value of $\alpha = 0.05$, and we use the standard normal distribution as our test statistic for testing the population parameter μ, then as we have discussed before, the difference between the assumption of null hypothesis, assumed value of this population parameter and the value obtained by the analysis of sample results is not expected to be more than $\pm 1.96\sigma_{\bar{X}}$ at $\alpha = 0.05$. This relationship can be shown by the following Figure 9.1.

NOTES

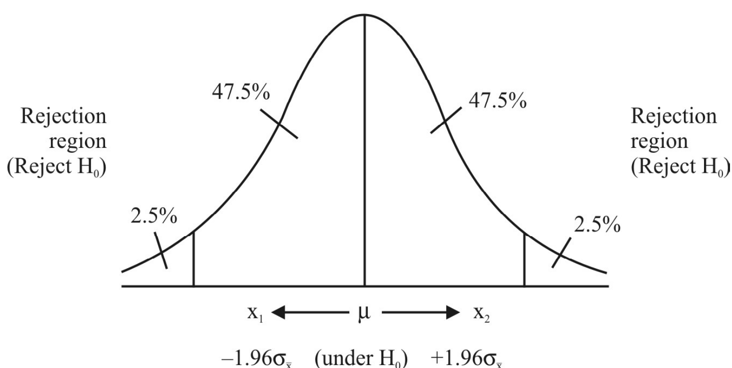


Fig. 9.1

In the above figure, if the sample $\overline{X}$ statistic falls within $1.96\,\sigma_{\bar{X}}$ of the assumed value of μ under the assumption of null hypothesis H_0, then we accept the null hypothesis as being correct at 95% confidence level (or 0.05 level of significance). The difference between $\bar{X}$ and μ which may be any value between X_1 and μ or X_2 and μ is considered to be accidental or due to chance element and is not considered significant enough or real enough to reject null hypothesis, so that for all practical purposes the value of $\bar{X}$ is considered equal to μ even though $\bar{X}$ can have any value between X_1 and X_2 as shown above. However, if the value of $\bar{X}$ falls beyond X_2 on the upper side or beyond X_1 on the lower side, then this difference between the values of $\bar{X}$ and μ would be considered significant and it will lead to rejection of null hypothesis. Since 5% of the time, this difference between the values of $\bar{X}$ and μ would be significant with 2.5% of the time $\bar{X}$ being too far above μ (beyond X_2) and 2.5% of the time being too far below μ (below X_1), the area of rejection will be on both sides of the mean extending into the tail sections of the curve. This area of rejection is known as the *critical region.*

5. ***Data collection and sample analysis.*** This involves the actual collection and computation of the sample data. A sample of the pre-established size *n* is collected and the estimate of the population parameter is calculated. This estimate is the value of the test statistic. For example, if we are testing a hypothesis about the value of population mean μ, then the test statistic would be the sample mean $\bar{X}$. Then we test this statistic to check whether it falls in the critical region or in the acceptance region. For example, if we want to test for the average IQ of the college students to be 130, then in that case we have to see that our population mean μ must be tested. We take a random sample of a given size *n* and calculate its mean $\bar{X}$ and then test it to see if the value of this $\bar{X}$ falls in the area of acceptance or in the area of rejection at a given level of significance.
6. ***Making the decision.*** Before the statistical decision is made, a decision rule must be established. Such decision rule will form the basis on which the null hypothesis will be accepted or rejected. This decision rule is really a formal

statement of the obvious purpose of the test. For example, this rule could be stated as follows,

Accept the null hypothesis if the value of sample statistic $\bar{X}$ falls within the area of acceptance, otherwise reject the null hypothesis.

NOTES

Based upon this established decision rule, a decision can be made whether to accept or reject the null hypothesis.

Committing Errors: Type I and Type II

- **Types of Errors**: There are two types of errors in statistical hypothesis, which are as follows:
  - o **Type I Error**: In this type of error, you may reject a null hypothesis when it is true. It means rejection of a hypothesis, which should have been accepted. It is denoted by α (alpha), and is also known alpha error.
  - o **Type II Error**: In this type of error, you are supposed to accept a null hypothesis when it is not true. It means accepting a hypothesis, which should have been rejected. It is denoted by β (beta), and is also known as beta error.

Type I error can be controlled by fixing it at a lower level, for example, If you fix it at 2%, then the maximum probability to commit Type I error is 0.02. But reducing Type I error, has a disadvantage when the sample size is fixed as it increases the chances of Type II error. In other words, it can be said that both types of errors cannot be reduced simultaneously. The only solution of this problem is to set an appropriate level by considering the costs and penalties attached to them or to strike a proper balance between both types of errors.

In a hypothesis test, a type I error occurs when the null hypothesis is rejected when it is in fact true; that is, H_0 is wrongly rejected. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; that is H_0: there is no difference between the two drugs on average. A type I error would occur if we concluded that the two drugs produced different effects when in fact there was no difference between them.

In a hypothesis test, a type II error occurs when the null hypothesis H_0, is not rejected when it is in fact false. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; that is H_0: there is no difference between the two drugs on average. A type II error would occur if it were concluded that the two drugs produced the same effect, that is, there is no difference between the two drugs on average, when in fact, they produced different ones.

In how many ways can we commit errors?

We reject a hypothesis when it may be true. This is Type I Error.

We accept a hypothesis when it may be false. This is Type II Error.

The other true situations are desirable:

We accept a hypothesis when it is true. We reject a hypothesis when it is false.

NOTES

	Accept H_0	**Reject H_0**
H_0 True	Accept True H_0 Desirable	Reject True H_0 Type I Error
H_1 False	Accept False H_0 Type II Error	Reject False H_0 Desirable

The level of significance implies the probability of type I error. A five per cent level implies that the probability of committing a type I error is 0.05. A one per cent level implies 0.01 probability of committing type I error.

Lowering the significance level and hence the probability of type I error is good but unfortunately it would lead to the undesirable situation of committing type II error.

To sum up:

- **Type I Error**: Rejecting H_0 when H_0 is true.
- **Type II Error**: Accepting H_0 when H_0 is false.

Note: The probability of making a Type I error is the level of significance of a statistical test. It is denoted by α.

Where, α = Prob. (Rejecting H_0 / H_0 true)

$1 - \alpha$ = Prob. (Accepting H_0 / H_0 true)

The probability of making a Type II error is denoted by β.

Where, β = Prob. (Accepting H_0 / H_0 false)

$1 - \beta$ = Prob. (Rejecting H_0 / H_0 false) = Prob. (The test correctly rejects H_0 when H_0 is false)

$1 - \beta$ is called the power of the test. It depends on the level of significance α, sample size n and the parameter value.

Null and Alternative Hypotheses

Hypothesis is usually considered as the principal instrument in research. The basic concepts regarding the testability of a hypothesis are as follows:

Null Hypothesis and Alternative Hypothesis

In the context of statistical analysis, while comparing any two methods, the following

NOTES

concepts or assumptions are taken into consideration:

- **Null Hypothesis**: While comparing two different methods in terms of their superiority, wherein the assumption is that both the methods are equally good is called null hypothesis. It is also known as statistical hypothesis and is symbolized as H_0.
- **Alternate Hypothesis**: While comparing two different methods, regarding their superiority, wherein, stating a particular method to be good or bad as compared to the other one is called alternate hypothesis. It is symbolized as H_1.

Comparison of Null Hypothesis with Alternate Hypothesis

Following are the points of comparison between null hypothesis and alternate hypothesis:

- Null hypothesis is always specific while alternate hypothesis gives an approximate value.
- The rejection of null hypothesis involves great risk, which is not in the case of alternate hypothesis.

Null hypothesis is more frequently used in statistics than alternate hypothesis because it is specific and is not based on probabilities.

The hypothesis to be tested is called the Null Hypothesis and is denoted by H_0. This is to be tested against other possible states of nature called alternative hypothesis. The alternative is usually denoted by $H_{1.}$

The null hypothesis implies that there is no difference between the statistic and the population parameter. To test whether there is no difference between the sample mean X and the population m, we write the null hypothesis.

$$H_0\text{: } \overline{X} = \mu$$

The alternative hypothesis would be,

$$H_1\text{: } \neq \mu$$

This means $> \mu$ or $< \mu$. This is called a two-tailed hypothesis.

The alternative hypothesis H_1: $> \mu$ is right tailed.

The alternative hypothesis H_1: $< \mu$ is left tailed.

These are one sided or one-tailed alternatives.

Note 1: The alternative hypothesis H_1 implies all such values of the parameter, which are not specified by the null hypothesis H_0.

Note 2: Testing a statistical hypothesis is a rule, which leads to a decision to accept or reject a hypothesis.

A one-tailed test requires rejection of the null hypothesis when the sample statistic is greater than the population value or less than the population value at a certain level of significance.

NOTES

1. We may want to test if the sample mean exceeds the population mean m. Then the null hypothesis is,

   H_0: > μ

2. In the other case the null hypothesis could be,

   H_0: < μ

Each of these two situations leads to a one-tailed test and has to be dealt with in the same manner as the two-tailed test. Here the critical rejection is on one side only, right for > μ and left for < μ. Both the Figures 9.2 and 9.3 here show a five per cent level of test of significance.

For example, a minister in a certain government has an average life of 11 months without being involved in a scam. A new party claims to provide ministers with an average life of more than 11 months without scam. We would like to test if, on the average, the new ministers last longer than 11 months. We may write the null hypothesis H_0: = 11 and alternative hypothesis H_1: > 11.

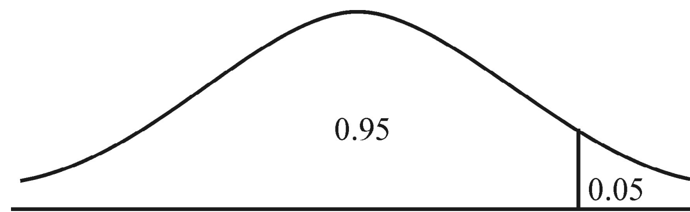


Fig. 9.2 H_0: $\bar{X}$ > μ

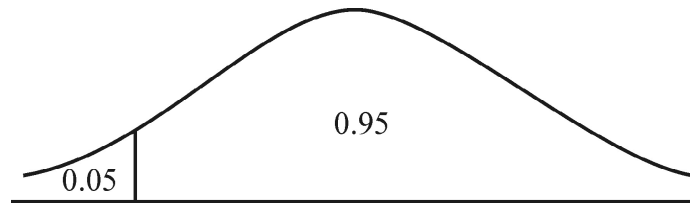


Fig. 9.3 H_0: < μ

9.4 TESTING THE OVERALL SIGNIFICANCE OF THE SAMPLE REGRESSION

So far we have discussed situations in which the null hypothesis is rejected if the sample statistic $\bar{X}$ is either too far above or too far below the population parameter μ, which means that the area of rejection is at both ends (or tails) of the normal curve. For example, if we are testing for the average IQ of the college students being equal to 130, then the null hypothesis $H_0 : \mu = 130$ will be rejected if a sample selected gives a mean $\bar{X}$ which is either too high or too low compared to μ. This can be expressed as follows:

$$H_0 : \mu = 130$$

$$H_1 : \mu \neq 130$$

This means that with $\alpha = 0.05$ (95% confidence interval), the value of $\bar{X}$ must be within $\pm\ 1.96\ \sigma_{\bar{X}}$ of the assumed value of μ under H_0 in order to accept the null hypothesis. In other words, $\frac{\bar{X} - \mu\,(\text{under } H_0)}{\sigma_{\bar{X}}}$ must be less than $\pm$ 1.96.

NOTES

The element $\frac{\bar{X} - \mu\,(\text{under } H_0)}{\sigma_{\bar{X}}}$ is known as the critical ratio or CR. It means that:

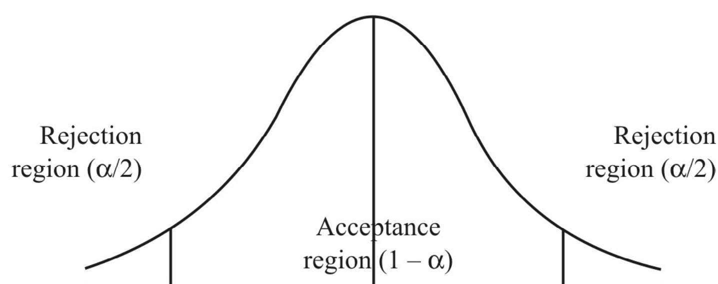
At α = 0.05, accept H_0 if critical ratio CR falls within ± 1.96 and reject H_0 if CR is less than (–1.96) or greater than (+1.96). If it happens to be exactly 1.96 then we can accept H_0.

On the other hand, there are situations in which the area of rejection lies entirely on one extreme of the curve, which is either the right end of the tail or the left end of the tail. Tests concerning such situations are known as *one-tailed* tests, and the null hypothesis is rejected only if the value of the sample statistic falls into this single rejection region.

For example, let us assume that we are manufacturing 9 volt batteries and we claim that our batteries last on an average (μ) 100 hours. If somebody wants to test the accuracy of our claim, he can take a random sample of our batteries and find the average ($\bar{X}$) of this sample. He will reject our claim only if the value of $\bar{X}$ so calculated is considerably lower than 100 hours, but will not reject our claim if the value of $\bar{X}$ is considerably higher than 100 hours. Hence in this case, the rejection area will only be on the left end tail of the curve.

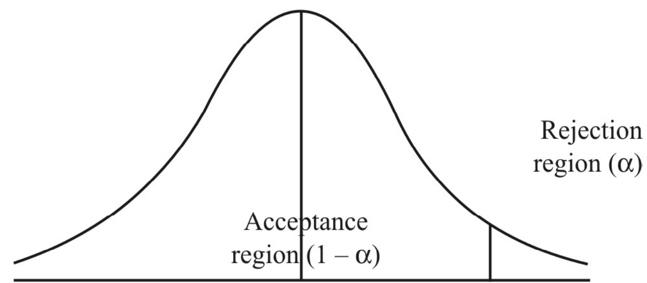
Similarly, if we are making a low calorie diet ice cream and claim that it has on an average only 500 calories per pound and an investigator wants to test our claim, he can take a sample and compute $\bar{X}$. If the value of $\bar{X}$ is much higher than 500 calories, then he will reject our claim. But he will not reject our claim if the value of $\bar{X}$ is much lower than 500 calories. Hence the rejection region in this case will be only on the right end tail of the curve. These rejection and acceptance areas are shown in the normal curves in Figure 9.4 as follows:

(*a*) *Two-Tailed Test*



(*b*) *Right-Tailed Test*

NOTES



(*c*) *Left-Tailed Test*

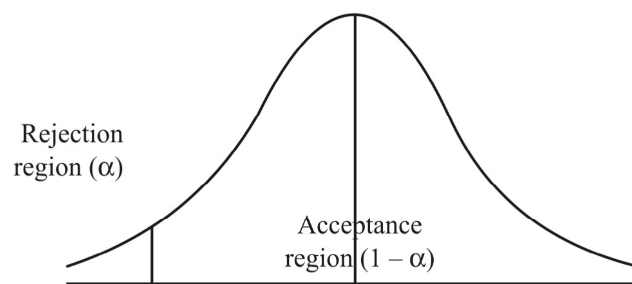


Fig. 9.4

Tests Involving a Population Mean (Large Sample)

This type of testing involves decisions to check whether a reported population mean is reasonable or not, compared to the sample mean computed from the sample taken from the same population. A random sample is taken from the population and its statistic $\bar{X}$ is computed. An assumption is made about the population mean μ as being equal to the sample mean and a test is conducted to see if the difference ($\bar{X} - \mu$) is significant or not. This difference is not significant if it falls within the acceptance region and this difference is considered significant if it falls within the rejection region or the critical region at a given level of significance α.

It must also be noted that if population is not known to be normally distributed, then the sample size should be large enough, generally more than 30. However, if population is known to be normally distributed and the population standard deviation is known then even a smaller sample size would be acceptable.

Example 9.1: (Two-Tailed Test)

Assume that the average annual income for government employees in the nation is reported by the Census Bureau to be $18,750.00. There was some doubt whether the average yearly income of government employees in Washington was representative of the national average.

A random sample of 100 government employees in Washington was taken and it was found that their average salary was $19,240.00 with a standard deviation of $2,610.00. At a level of significance $\alpha = 0.05$ (95% confidence level),

NOTES

can we conclude that the average salary of government employees in Washington is representative of the national average?

Solution: Obviously, it is a two-tailed test because if the salary of government employees in Washington is too high or too low compared to the national average then the hypothesis that the average salary of government employees in Washington is no different than the national average would be rejected.

Following the steps described in the procedure for hypothesis testing, we find:

1. Null hypothesis: $H_0 : \mu = \$18,750$

   Alternate hypothesis: $H_1 : \mu \neq \$18,750$
2. Level of significance as given $\alpha = 0.05$.
3. Determination of a suitable test statistic. Since we are testing for the population mean and according to the Central Limit Theorem, the sampling distribution of the sample means is approximately normally distributed with a standard error of the mean being $\sigma_{\bar{X}}$, the following test statistic would be appropriate:

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \quad \textit{where } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}},$$

Hence,
$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Where,

$\bar{X}$ = Sample mean

μ = Population mean

σ = Standard deviation of the population (s)

(Since population standard deviation is not given and the sample size is large enough, we can approximate the sample standard deviation s as equivalent to population standard deviation σ.)

4. Defining the critical region. Since $\alpha = 0.05$ and it is a two-tailed test, the rejection region will be on both end tails of the curve in such a way that the rejection area will comprise 2.5% at the end of the right tail and 2.5% at the end of the left tail. In other words, at $\alpha = 0.05$, the region of acceptance is enclosed by the value of Z being ± 1.96 around the mean.

Now for our example, let us calculate the value of Z.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Where,

$\bar{X}$ = 19240

μ = 18750

$\sigma = s = 2610$

$n = 100$

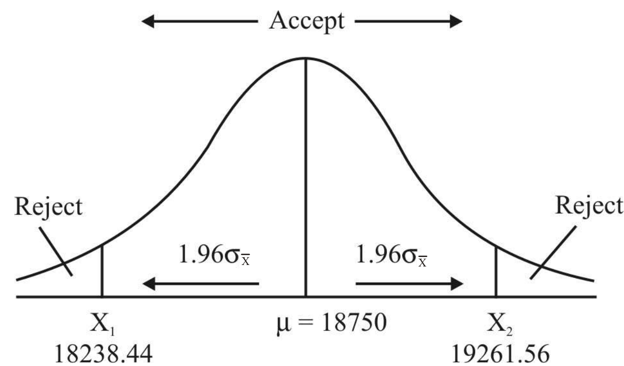
NOTES

Then, $$Z = \frac{19240 - 18750}{2610/\sqrt{100}}$$

$$= \frac{490}{261} = 1.877$$

Now, since the calculated value of Z as 1.877 is less than 1.96 and falls within the area of acceptance bounded by Z = ±1.96, we cannot reject the null hypothesis.

We could also solve this problem by constructing a 95% confidence interval for the population mean and then testing whether the sample mean falls within the confidence interval. The confidence interval is bounded by $\mu \pm 1.96\ \sigma_{\bar{X}}$, as illustrated below:



Now, $X_1 = \mu - Z\sigma_{\bar{X}}$

and $X_2 = \mu + Z\sigma_{\bar{X}}$

We know that, $\mu = 18750$

$Z = 1.96$ at $\alpha = 0.05$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}} = \frac{2610}{\sqrt{100}} = 261$$

Then, $X_1 = 18750 - 1.96(261)$

$= 18238.44$

and $X_2 = 18750 + 1.96(261)$

$= 19261.56$

This means that if the sample mean lies within these two limits, then we cannot reject the null hypothesis. As we can see, the sample mean of \$19,240.00 lies within this interval, so that we cannot reject the null hypothesis. Hence, our

decision is to conclude that there is no significant difference between the average salary of government employees in Washington and the national average and it is purely coincidental that the average salary of government employees in Washington is numerically different than the national average.

Example 9.2: One-Tailed Test (Left-Tail)

The manufacturer of light bulbs claims that a light bulb lasts on an average 1600 hours. We want to test his claim. We will not reject his claim if the average of the sample taken lasts considerably more than 1600 hours, built we will reject his claim if it lasts considerably less than 1600 hours. Hence, it is a one-tailed test and the area of rejection is the left end tail of the curve.

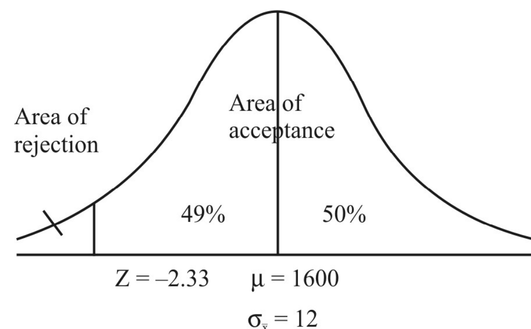
A sample of 100 light bulbs was taken at random and the average bulb life of this sample was computed to be 1570 hours with a standard deviation of 120 hours. At $\alpha = 0.01$, let us test the validity of the claim of this manufacturer.

Solution: Since the sample is large ($n = 100$), we can approximate the population standard deviation (σ) by sample standard deviation (s) so that:

Null hypothesis: $H_0 : \mu = 1600$

Alternate hypothesis : $H_1 : \mu < 1600$

Then at 99% confidence interval ($\alpha = 0.01$), the acceptance region is bounded by Z = –2.33 on the left tail of the standardized normal curve as shown below:



Now, $$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

Where, $\bar{X} = 1570$

$\mu = 1600$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}} = \frac{120}{\sqrt{100}} = 12$$

Then, $$Z = \frac{1570 - 1600}{12} = -\left(\frac{30}{12}\right) = -(2.5)$$

NOTES

Since our computed value of Z is numerically larger than the critical value of Z which is – (2.33), we cannot accept the null hypothesis at 99% confidence interval. (The negative sign is simply a concept that the value lies on the left of the mean μ and it is not an algebraic sign.) This means that the manufacturer's claim is not valid.

Example 9.3: One-Tailed Test (Right Tail)

An insurance company claims that it takes 2 weeks (14 days), on an average, to process an auto accident claim. The standard deviation is 6 days. To test the validity of this claim, an investigator randomly selected 36 people who recently filed claims. This sample revealed that it took the company an average of 16 days to process these claims. At 99% level of confidence, check if it takes the company more than 14 days on an average to process a claim.

Solution: In this case, the population parameter being tested is μ which is the average number of days it takes the company to process a claim. The company's claim is not valid if it takes considerably longer than the 14 days it claims on an average to process a claim. Hence,

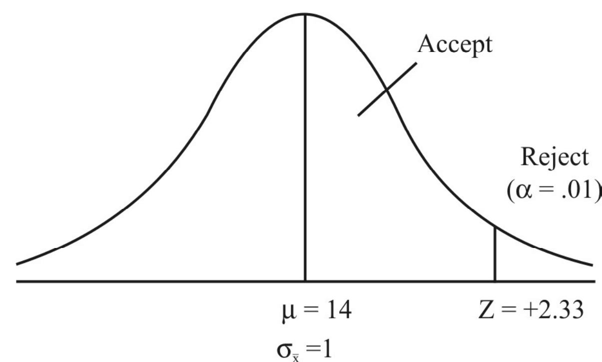
$$H_0 : \mu = 14$$

$$H_1 : \mu > 14$$

Then,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$$Z = \frac{16 - 14}{6/\sqrt{36}} = \frac{2}{1} = 2$$



Since the Z value of 2 is less than the critical value of Z which is 2.33, and it falls within the region of acceptance, we cannot reject the null hypothesis. Accordingly the company's claim is considered to be valid.

Tests Involving A Single Proportion

So far, we have dealt with the population parameter μ which reflects quantitative data. It cannot be used for qualitative data. For such qualitative data, the parameter

of interest is the population proportion favouring one of the outcomes of the event. There are many situations in which we must test the validity of statements about the population proportions or percentages. For example, if a politician claims that 60% of the population supports his viewpoint on a given issue, we can test this claim by taking random samples of people and asking their opinions about this politician and finding the percentage of people on an average who support the viewpoint of this politician and then test whether this sample percentage is significantly different than his claim of population percentage. This technique is used in analysing the qualitative data where we can test for the presence or absence of a certain characteristic. For example, we may want to know if the government figures on the unemployment situation are accurate or not. Suppose that the government figures indicate that 9% of the work force is unemployed. We can always take a random sample and check for its validity.

This type of data follows the binomial distribution with:

Sample proportion $p = \frac{x}{n}$

However, if n is large enough, so that,

$$np \geq 5$$

and $$n(1-p) \geq 5$$

Then it can be approximated to normal distribution and test statistic Z can be used. Where,

$$Z = \frac{p - \pi}{\sigma_p}$$

π = Population proportion

p = Sample proportion

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} \text{ or } = \sqrt{\frac{p(1-p)}{n}}$$

Then the computed value of Z is compared with the critical value of Z in order to accept or reject the null hypothesis.

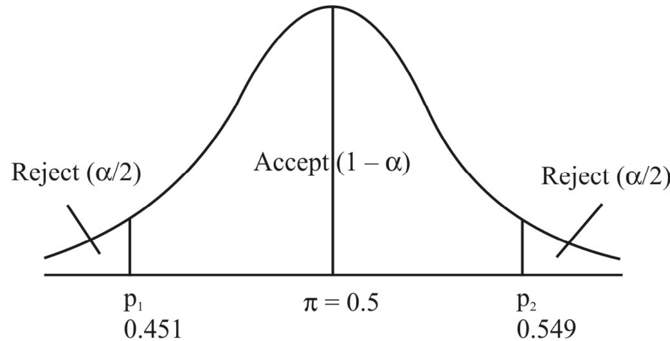
The testing of hypothesis follows the same procedure as in the case of tests about the population means and can best be illustrated with the help of following example.

Example 9.4: The sponsor of a television show believes that his studio audience is divided equaly between men and women. Out of 400 persons attending the show one day, there where 230 men. At $\alpha = 0.05$, test if the belief of the sponsor is correct.

Solution: This is a two-tailed test, since too many men as well as too few men in the audience would become the cause of rejection of the null hypothesis.

In order for the null hypothesis to be accepted, the sample proportion $p = (x/n)$ must fall within the confidence interval bounded by p_1 and p_2 as shown in the diagram which is the area of acceptance.

NOTES



Here,

Null hypothesis: $H_0 : p = 0.5$

Alternate hypothesis: $H_1 : p \neq 0.5$

Confidence interval is defined as follows:

$$p_1 = p - Z\sigma_p$$
$$p_2 = p + Z\sigma_p$$

Where,

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{(0.5)(0.5)}{400}} = 0.025$$

$Z = 1.96$ (at $\alpha = 0.05$)

$\pi = 0.5$

Substituting these values we get,

$$p_1 = 0.5 - 1.96(0.025) = 0.451$$
$$p_2 = 0.5 + 1.96(0.025) = 0.549$$

In our example, the sample proportion $p = x/n = 230/400 = 0.575$. Clearly our sample proportion lies outside the region of acceptance and is in the critical region. Hence the null hypothesis cannot be accepted.

An alternate method to test the validity of the null hypothesis would be to compute the value of Z for the given information and compare it with the critical value of Z from the table, which, at 0.05 level of significance is 1.96.

Now,

$$Z = \frac{p - \pi}{\sigma_p}$$

Where,

p = Sample proportion = 230/400 = 0.575

π = Population proportion = 0.5

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = 0.025 \text{ (as calculated above).}$$

NOTES

Then, $$Z = \frac{0.575 - 0.500}{0.025}$$

$$= \frac{0.075}{0.025} = 3$$

Since our computed value of Z = 3, is higher than the critical value of Z = 1.96, we cannot accept the null hypothesis.

Example 9.5: One-Tailed Test

The mayor of the city claims that 60% of the people of the city follow him and support his policies. We want to test whether his claim is valid or not. A random sample of 400 persons was taken and it was found that 220 of these people supported the mayor. At level of significance $\alpha = 0.01$ what can we conclude about the mayor's claim.

Solution: Clearly, it is a one-tailed test for we will only reject the mayor's claim if the sample proportion of persons who support the mayor is considerably less than the mayor's claim about the population proportion of persons who support him. We will not reject his claim if such sample proportion is considerably higher than the population proportion.

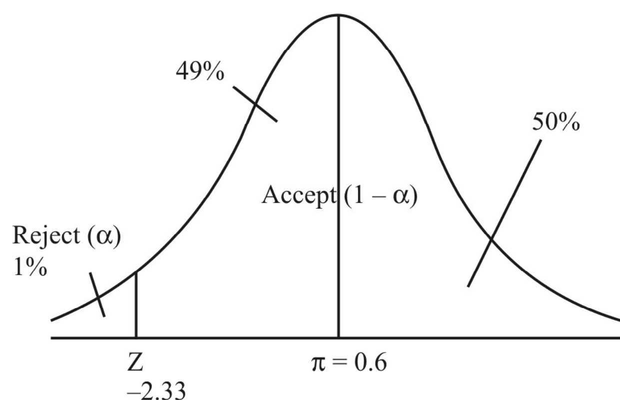
Hence:

Null hypothesis: $H_0 : \pi = 0.6$

Alternate hypothesis: $H_1 : \pi < 0.6$

(The null hypothesis may also be expressed as $H_0 : \pi \geq 0.6$).

The sample proportion $p = 220/400 = 0.55$



NOTES

Now,

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

$$= \sqrt{\frac{(0.6)(0.4)}{400}}$$

$$= \sqrt{0.006} = 0.0245$$

Then,

$$Z = \frac{p-\pi}{\sigma_p}$$

$$= \frac{0.55-0.6}{0.0245}$$

$$= -\left(\frac{0.05}{0.0245}\right) = -(2.04)$$

Since, it is a one-tailed test, the critical value of Z = – (2.33) for $\alpha = 0.1$. Ignoring the negative sign, we note that the numerical value of our computed Z is less than the numerical critical value of Z, and hence we cannot reject the null hypothesis.

Two Sample Test for Large Samples

In many decision-making situations, comparison of two population means or two population proportions, becomes an area of interest. For example, we may be interested in comparing the effectiveness of two different teaching methods, where the effectiveness would be measured by the difference in the average student achievement under the two different techniques. Or, we may be interested to know if there is any significant difference in the average age of life for men and women in this country. Or, we may be interested to know if the average expenditure of two different communities are significantly different from each other. For this purpose, we can test one population mean against the other and draw conclusions for the purpose of making rational decisions.

Testing the Difference between Two Sample Means

So far we have discussed sampling distribution of the means where a hypothesis was tested for any significant difference between the sample mean and the population mean. Now, we are interested to know if there are any significant differences between two population means. Let us assume that we want to find out if there is any significant difference in the average age of students who graduate with a bachelor degree in business from Baruch college and from Medgar Evers college. We take corresponding samples of graduating seniors from both colleges and find the mean

NOTES

of each sample taken from each college. Let these means and the differences in these means be represented as follows:

Baruch(1)	*Medgar Evers (2)*	*Differences*
$\bar{X}_{11}$	$\bar{X}_{21}$	$\bar{X}_{11} - \bar{X}_{21}$
$\bar{X}_{12}$	$\bar{X}_{22}$	$\bar{X}_{12} - \bar{X}_{22}$
$\bar{X}_{13}$	$\bar{X}_{23}$	$\bar{X}_{13} - \bar{X}_{23}$
$\bar{X}_{14}$	$\bar{X}_{24}$	$\bar{X}_{14} - \bar{X}_{24}$
•	•	•
•	•	•
•	•	•
$\bar{X}_{1n}$	$\bar{X}_{2n}$	$\bar{X}_{1n} - \bar{X}_{2n}$

In the above example, the first subscript represents the college and the second subscript represents the sequential sample number.

Now we have a distribution of the differences in the sample means. This is known as the sampling distribution of $(\bar{X}_1 - \bar{X}_2)$.

Basing on our analysis of the Central Limit Theore, we can make the following statements concerning the sampling distribution of the difference between sample means $(\bar{X}_1 - \bar{X}_2)$.

If two independent samples of size n_1 and n_2 (both n_1 and n_2 to be larger than 30) are taken from populations with mean μ_1 and μ_2, and standard deviation σ_1 and σ_2, distribution with the following properties,

(a) The mean of the sampling distribution of $(\bar{X}_1 - \bar{X}_2)$ is $(\mu_1 - \mu_2)$.

(b) The standard error of differences of sample means $\sigma_{(\bar{X}_1 - \bar{X}_2)}$ is given by,

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

However, if σ_1 and σ_2 are not known, then since n_1 and n_2 are sufficiently l;arge, the standard error of this distribution can be approximated by,

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

For the purpose of testing the hypothesis, we can use the standard normal distribution to find the Z score as,

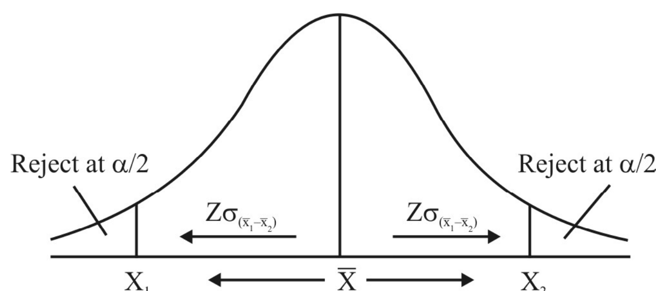
$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

or, $$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

NOTES

Then the decisions can be made on the basis of whether the value of Z so calculated falls within the region of acceptance or whether it falls in the region of rejection at a given value of the level of significance.

Another way to test for the significance of such a difference is to put one sample mean as the mean of the normal distribution and see if the second sample mean lies within the region of acceptance or not, at a given value of α. If the second sample mean lies within the acceptance region (within the bounds of X_1 and X_2 as shown below) then we can accept the null hypothesis that there is no significant difference between the two population means and that both samples come from the same population and any numerical difference in values of these two sample means happened by chance or due to a sampling error.



Example 9.6: A potential buyer of electric bulbs bought 100 bulbs each of two famous brands, A and B. Upon testing both these samples, he found that brand A had a mean life of 1500 hours with a standard deviation of 50 hours whereas brand B had an average life of 1530 hours with a standard deviation of 60 hours. Can it be concluded at 5% level of significance ($\alpha = 0.05$) that the two brands differ significantly in quality?

Solution: We assume that there is no significant difference in the quality of both brands so that brand A is as good as brand B in terms of average number of operating hours, so that,

Null hypothesis: $H_0 : \mu_1 = \mu_2$

Alternate hypothesis: $H_1 : \mu_1 \neq \mu_2$

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where, $$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

NOTES

$$= \sqrt{\frac{(50)^2}{100} + \frac{(60)^2}{100}}$$

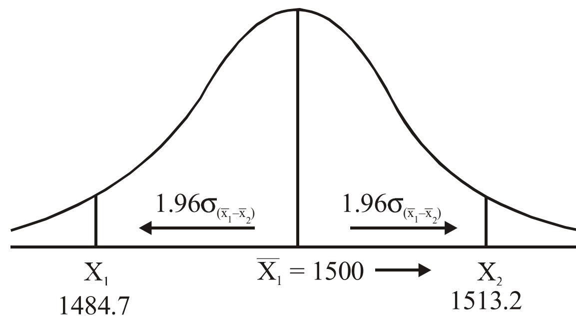
$$= \sqrt{61} = 7.81$$

Hence, $$Z = \frac{1500 - 1530}{7.81}$$

$$= -\left(\frac{30}{7.81}\right) = -(3.841)$$

Since the computed numerical absolute value of Z is more than the critical value of Z from the table at $\alpha = 0.05$, which is = 1.96, we cannot accept the null hypothesis.

We could also solve this problem by establishing the confidence interval where interval boundaries are as given:



In our case, we know that:

$$\sigma_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 7.81$$

Then,

$$\text{Lower limit} = X_1 = \bar{X} - Z\sigma_{(\bar{X}_1 - \bar{X}_2)}$$

$$= 1500 - 1.96(7.81) = 1484.7$$

$$\text{Upper limit} = X_1 = \bar{X} - Z\sigma_{(\bar{X}_1 - \bar{X}_2)}$$

$$= 1500 + 1.96(7.81) = 1515.3$$

Since the value of $\bar{X}_2$ as 1530 lies beyond the acceptable limit of 1515.3, we cannot accept the null hypothesis.

Even though, our example above is a two-tailed test, we can also perform one-tailed tests for the differences between two population means where the null hypothesis is rejected when one mean is significantly higher or significantly lower

than the other mean. These one-tailed tests are conceptually similar to the one-tailed tests of a single mean discussed earlier. We can illustrate this with the help of following example.

NOTES

Example 9.7: A civil group in the city claims that a female college graduate earns less than a male college graduate. To test this claim, a survey of starting salary of 60 male graduates and 50 female graduates was taken and it was found that the average starting salary for female graduates was $29,500 with a standard deviation of $500 and the average salary for male graduates was $30,000 with a standard deviation of $600. At 1% level of significance, ($\alpha = 0.01$), test if the claim of this civil group is valid.

Solution: The civil group claims that the average starting salary of female graduates is considerably less than the average starting salary of male graduates. The null hypothesis states that the starting salary of female graduates is not less than the starting salary of male graduates. Accordingly the null hypothesis will be rejected only if the average starting salary of female graduates is significantly less than the corresponding average starting salary of male graduates. The null hypothesis will not be rejected if this average is considerably higher than the average starting salary of male graduates. Hence, it is a one-tailed test.

Let $\bar{X}_1$ and s_1 represent the sample mean and the standard deviation, respectively, of the starting salary of female graduates.

Similarly, let $\bar{X}_2$ and s_2 respectively represent the mean and the standard deviation of the starting salary of male graduates. This data can be represented as follows:

Females	*Males*
$\bar{X}_1 = \$29{,}500$	$\bar{X}_2 = \$30{,}000$
$s_1 = \$500$	$s_2 = \$600$
$n_1 = 50$	$n_2 = 60$

We have to test whether at $\alpha = 0.01$, the observed difference between $\bar{X}_1$ and $\bar{X}_2$ is significant or not.

Hence, $\quad H_0 : \mu_1 \geq \mu_2$

$H_1 : \mu_1 < \mu_2$

Now,
$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

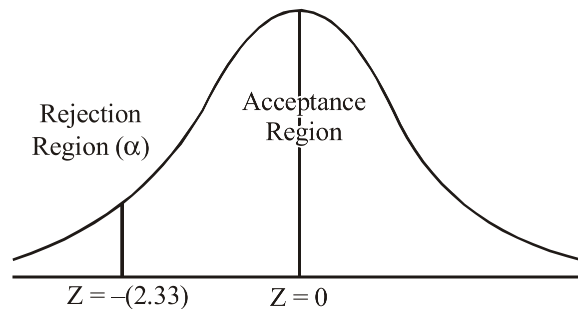
$$= \frac{29{,}500 - 30{,}000}{\sqrt{\dfrac{(500)^2}{50} + \dfrac{(600)^2}{60}}}$$

$$= -\left(\frac{500}{\sqrt{5000+6000}}\right)$$

$$= -\left(\frac{500}{104.88}\right) = -(4.77)$$

NOTES

At 1% level of significance, the critical value of Z on the left-end of the tail for a one-tailed test is –(2.33). Since our computed absolute numerical value of Z is higher than the absolute critical value of Z, we cannot accept the null hypothesis. This is illustrated in the following Z score normal distribution curve.



Testing for the Difference of Two Population Proportions

In some situations, it is necessary to check whether the two population proportions are equal or not. Suppose we want to check whether the percentage of female students entering college after completing high school is significantly different than the percentage of male students similarly entering college. Or suppose, that we want to test whether the proportion of people supporting a national political leader in the north of the country is similar to proportion of people supporting him in the south. These tests require comparisons of two proportions to see if any difference between them is significant or not.

Distribution of Differences in Proportions

Since we are trying to find out if the difference between two population proportions is significant or not, we need to know the distribution of *differences of sample proportions*, just as we did in the case of comparison of two sample means earlier. This concept can best be illustrated by an example.

Suppose that we select 10 random samples of 200 students each (n_1) at Medgar Evers college and record the proportion (p_1) of females in each sample. Similarly, we also select 10 samples of 200 students each (n_2) from Baruch College and record the proportion of females (p_2) for each sample. These proportions and their differences ($p_1 - p_2$) for each paired sample are tabulated below:

NOTES

Sample	Medgar Evers (p_1)	Baruch (p_2)	Difference ($p_1 - p_2$)
1	0.64	0.57	0.07
2	0.65	0.60	0.05
3	0.58	0.58	0.00
4	0.62	0.65	–0.03
5	0.56	0.62	–0.06
6	0.66	0.61	0.05
7	0.60	0.55	0.05
8	0.59	0.59	0.00
9	0.62	0.57	0.05
10	0.58	0.56	0.02

The distribution of values of $(p_1 - p_2)$ above is known as the distribution of *differences of sample proportions*. Theoretically, if we took all possible pairs of random samples from these two populations and found the proportion of females in these samples and calculated the differences in each sample $(p_1 - p_2)$, then the resulting distribution of these differences will be approximately normally distributed with the following characteristics:

1. Since these proportions are represented by binomial distribution and we are approximating the binomial distribution to the normal distribution, the sample sizes from each population should be large enough. In general,

$$n_1 p_1 \geq 5$$
$$n_2 p_2 \geq 5$$
$$n_1 q_1 \geq 5$$
$$n_2 q_2 \geq 5$$

2. The mean of the distribution of differences of proportions is given by $(\pi_1 - \pi_2)$, where *px* equals the proportion of female students in the population of all students at Medgar Evers college and π_2 is the proportion of female students in the population of all students at Baruch college.
3. The standard deviation of the distribution of differences in proportions is denoted by $\hat{\sigma}_p$ (sigma sub *p* hat) and is given by:

$$\hat{\sigma}_p = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Where $\hat{\pi}$ (pi hat) is the pooled estimate of the values p_1 and p_2 under null hypothesis, which assumes that there is no difference between the two population proportions. This $\hat{\pi}$ is given by,

$$\hat{\pi} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Then using the test for Z scores, since it is approximated as normal distribution, we get,

$$Z = \frac{p_1 - p_2}{\hat{\sigma}_p}$$

We compare this computed value of Z with the critical value of Z from the table under a given level of significance and decide whether to accept or reject the null hypothesis.

Two-Tailed Test for Differences between Two Proportions

Example 9.8: A sample of 200 students at Baruch college revealed that 18% of them were seniors. A similar sample of 400 students at Hunter college revealed that 15% of them were seniors. To test whether the difference between these two proportions is significant enough to conclude that these populations are indeed different at 5% level of significance ($\alpha = 0.05$).

Solution: Null hypothesis: $H_0 : \pi_1 = \pi_2$

Alternate hypothesis: $H_1 : \pi_1 \neq \pi_2$

It is a two-tailed test because if the proportion of seniors at Baruch college is significantly higher than the proportion of seniors at Hunter college, then the null hypothesis will be rejected and similarly, if the proportion of seniors at Baruch college is significantly lower than the proportion of seniors at Hunter college, the null hypothesis will again be rejected.

Now,

The proportion of seniors at Baruch college = $p_1 = 0.18$

The proportion of seniors at Hunter college = $p_2 = 0.15$

Then,

$$Z = \frac{p_1 - p_2}{\hat{\sigma}_p}$$

Let us first calculate the value of $\hat{\sigma}_p$. We know that,

$$\hat{\sigma}_p = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

and,

$$\hat{\pi} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Here,

$n_1 = 200$

$n_2 = 400$

$p_1 = 0.18$

$p_2 = 0.15$

Substituting these values, we get,

NOTES

$$\hat{\pi} = \frac{200(0.18)+400(0.15)}{200+400}$$

$$= \frac{36+60}{600} = \frac{96}{600} = 0.16$$

Substituting the value of $\hat{\pi}$ we calculate the value of $\hat{\sigma}_p$ as,

$$\hat{\sigma}_p = \sqrt{(0.16)\ (0.84)\left(\frac{1}{200}+\frac{1}{400}\right)}$$

$$= \sqrt{0.1344 \times 0.0075} = \sqrt{0.001008} = 0.0317$$

Now,

$$Z = \frac{p_1 - p_2}{\hat{\sigma}_p}$$

$$= \frac{0.18-0.15}{0.0317} = \frac{0.03}{0.0317} = 0.95$$

Since our computed value of Z is less than the critical value of Z = 1.96 at $\alpha = 0.05$, for a two-tailed test, we cannot reject the null hypothesis.

One-Tailed Test for Difference between Two Proportions

Conceptually, the one-tailed test for differences between two population proportions is similar to a one-tailed test for the difference between two population means and the area of rejection will lie only in one end of the normal curve, either in the left end tail or in the right end tail, depending upon the type of problem.

Example 9.9: An insurance company believes that smokers have higher incidence of heart diseases than non-smokers in men over 50 years of age. Accordingly, it is considering to offer discounts on its life insurance policies to non-smokers. However, before the decision can be made, an analysis is undertaken to justify its claim that the smokers are at a higher risk of heart disease than non-smokers. The company randomly selected 200 men which included 80 smokers and 120 non-smokers. The survey indicated that 18 smokers suffered from heart disease and 15 non-smokers suffered from heart disease. At 5% level of significance, can we justify the claim of the insurance company that smokers have a higher incidence of heart disease than non-smokers?

Solution: Let p_1 be the proportion of male smokers over 50 years of age who suffer from heart disease in the entire population and let p_2 be the corresponding proportion of non-smokers. Then,

NOTES

$$\text{Null hypothesis: } H_0 : \pi_1 = \pi_2$$

$$\text{Alterante hypothesis: } H_1 : \pi_1 > \pi_2$$

$$\text{Test statistic: } Z = \frac{p_1 - p_2}{\hat{\sigma}_p}$$

Now,

p_1= Proportion of male smokers over 50 years of age who suffer from heart disease,

$$= \frac{18}{80} = 0.225$$

p_2= Proportion of male non-smokers over 50 years of age who suffer from heart disease,

$$= \frac{15}{120} = 0.125$$

$$\hat{\sigma}_p = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\hat{\pi} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$= \frac{80(0.225) + 120(0.125)}{80 + 120}$$

$$= \frac{18 + 15}{200}$$

$$= \frac{33}{200} = 0.165$$

Then,

$$\hat{\sigma}_p = \sqrt{(0.165)(0.835)\left(\frac{1}{80} + \frac{1}{120}\right)}$$

$$= \sqrt{(0.1378)(0.0208)}$$

$$= \sqrt{0.00287} = 0.0536$$

Hence,

$$Z = \frac{0.225 - 0.125}{0.0536}$$

$$= \frac{0.1}{0.0536} = 1.86$$

NOTES

Since the critical value of Z at $\alpha = 0.05$ for a one-tailed test is 1.64 and since our computed value of $Z = 1.86$ is higher than the critical value of Z, we cannot accept the null hypothesis. It shows that there is a strong evidence to infer that the proportion of smokers who have heart diseases is greater than the proportion of non-smokers who have heart disease.

Check Your Progress

1. Elaborate on the normality assumption.
2. What do you understand by the hypothesis?
3. Define the statistical hypothesis.
4. Explain the types of statistical hypothesis.
5. Illustrate the testing of simple hypothesis.
6. State the one-tailed or two-tailed hypothesis.
7. Interpret the critical region.
8. Elaborate on the size/significance level of a test.
9. Define the two sample test for large samples.

9.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. In multiple regression, the assumption needing a normal distribution applies only to the disturbance term, not to the independent variables as is often supposed.
2. A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. Hypothesis refers to a provisional idea whose merit needs evaluation, but having no specific meaning. Though it is often referred as a convenient mathematical approach for simplifying cumbersome calculation.
3. Statistical hypothesis cannot ascertain the truth of the population parameter. To do this, truth table for entire population is required to be examined which is time consuming and impractical. Researchers examine a random sample and after judging its consistency, the hypothesis is accepted.
4. Statistical hypotheses are of two types:
   - Null Hypothesis: It signifies that sample observations result purely from chance. Its notation is H_0.
   - Alternative Hypothesis: It signifies that sample observations are influenced by some cause of non-random nature. Its notation is H_1 or H_a.

5. Many applications in language engineering require testing of hypotheses. Suppose we have to differentiate between a person's 'Reading a Speech' and 'Giving Spontaneous Speech'.
6. The decision on selection of alternate hypotheses, to propose, depends on factors leading the language engineer to note differences in one direction or the other. These instances are referred to as one-tailed or two-tailed hypotheses depending on whether differentiation is being done for one direction or both the directions.
7. This region is defined as a set of test statistic leading to rejection of the null hypothesis in a hypothesis test. For this the sample space is split into two mutually exclusive regions. This region provides basis to reject the null hypothesis H_0.
8. Significance level in simple hypothesis test is the probability of incorrectly rejecting the null hypothesis. In a composite hypothesis it is the upper bound of the probability that serves the basis of rejecting the null hypothesis.
9. In many decision-making situations, comparison of two population means or two population proportions, becomes an area of interest. For example, we may be interested in comparing the effectiveness of two different teaching methods, where the effectiveness would be measured by the difference in the average student achievement under the two different techniques.

9.6 SUMMARY

- In multiple regression, the assumption needing a normal distribution applies only to the disturbance term, not to the independent variables as is often supposed.
- In actual, each case in the sample has a different random variable which encompasses all the "Noise" that accounts for differences in the observed and predicted values produced by a regression equation.
- In econometrics, normality tests are used to determine if a data set is well-modelled by a normal distribution and to compute how likely it is for a random variable underlying the data set to be normally distributed.
- In descriptive statistics terms, one measures a goodness of fit of a normal model to the data – if the fit is poor then the data are not well modelled in that respect by a normal distribution, without making a judgment on any underlying variable.
- A normality test is used to determine whether sample data has been drawn from a normally distributed population (within some tolerance). A number of statistical tests, such as the Student's t-test and the one-way and two-way ANOVA require a normally distributed sample population.

NOTES

- There are two approaches to statistical inference: model-based inference and design-based inference. Both approaches rely on some statistical model to represent the data-generating process.
- A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. Hypothesis refers to a provisional idea whose merit needs evaluation, but having no specific meaning. Though it is often referred as a convenient mathematical approach for simplifying cumbersome calculation.
- Hypothesis is put forward as a proposition. It may even be a set of more than one proposition. A proposition is the antecedent of a conditional proposition which may be an assumption or a guess. It is something yet to be proved, but taken to be temporarily true.
- Formalized hypotheses have two variables, independent and dependent. The independent variable is the person, may be the scientist, who is going to put the hypothesis and the dependent variable is one that the person observes.
- Statistical hypothesis cannot ascertain the truth of the population parameter. To do this, truth table for entire population is required to be examined which is time consuming and impractical. Researchers examine a random sample and after judging its consistency, the hypothesis is accepted.
- The decision on selection of alternate hypotheses, to propose, depends on factors leading the language engineer to note differences in one direction or the other. These instances are referred to as one-tailed or two-tailed hypotheses depending on whether differentiation is being done for one direction or both the directions.
- Significance level in simple hypothesis test is the probability of incorrectly rejecting the null hypothesis. In a composite hypothesis it is the upper bound of the probability that serves the basis of rejecting the null hypothesis.
- In many decision-making situations, comparison of two population means or two population proportions, becomes an area of interest. For example, we may be interested in comparing the effectiveness of two different teaching methods, where the effectiveness would be measured by the difference in the average student achievement under the two different techniques.

9.7 KEY WORDS

- **Normality assumption:** In multiple regression, the assumption needing a normal distribution applies only to the disturbance term, not to the independent variables as is often supposed.
- **Hypothesis:** A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. Hypothesis refers to a provisional idea whose merit needs evaluation, but having no specific meaning.

- **Statistical hypothesis:** Statistical Hypothesis cannot ascertain the truth of the population parameter. To do this, truth table for entire population is required to be examined which is time consuming and impractical.
- **Null hypothesis:** It signifies that sample observations result purely from chance. Its notation is H_0.
- **Alternative hypothesis:** It signifies that sample observations are influenced by some cause of non-random nature. Its notation is H_1 or H_a.
- **Critical region:** This region is defined as a set of test statistic leading to rejection of the null hypothesis in a hypothesis test. This region provides basis to reject the null hypothesis H_0.

9.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Explain the term normality assumption.
2. Elaborate on the hypothesis.
3. Illustrate the statistical hypothesis.
4. Define the types of statistical hypothesis.
5. Interpret the testing of simple hypothesis.
6. What do you mean by the one-tailed or two-tailed hypothesis?
7. State the critical region.
8. Define the size/significance level of a test.
9. Explain the two sample test for large samples.

Long-Answer Questions

1. Explain the normality assumption with the help of examples.
2. Briefly discuss the hypothesis testing about individual partial regression coefficients.
3. Compare the null hypothesis and alternate hypothesis.
4. Describe the testing the overall significance of the sample regression. Give appropriate examples.

9.9 FURTHER READINGS

Johnston, J. and John DiNARDO. 1997. *Econometric Methods*, Fourth Edition. New Delhi: Tata McGraw-Hill.

Koutsoyiannis, A.1977. *Theory of Econometrics,* Second Edition. London: The Macmillan Press Ltd.

NOTES

Özdemir, Durmuº. 2016. *Applied Statistics for Economics and Business,* Second Edition. Izmir (Turkey): Springer.

Maddala, G. S. 1992. *Introduction to Econometrics,* Second Edition. New York: Macmillan Publishing Company.

Pindyck, R. S and D. L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts,* Fourth Edition. New York: McGraw Hill.

Goldberger, A. S. 1998. *Introductory Econometrics*. Cambridge: Harvard University Press.

Levine, David M., Timothy C. Krehbiei, Mark L. Berenson and P. K. Viswanathan. 2009. *Business Statistics*, Fifth Edition. New Delhi: Pearson Education.

Webster, Allen L. 1998. *Applied Statistics for Business and Economics,* Third Edition. New Delhi: Tata McGraw-Hill.

UNIT 10 LINEAR RESTRICTIONS

NOTES

Structure

10.0 INTRODUCTION

In linear regression, linear represents the relationship between the parameters being estimated and the outcome. In a linear model, the estimate of the parameter vector can be written as ˆβ=Σiwiyi, where the {wi} are weights determined by the estimation procedure.

When conducting individual *t*–tests, a restriction is imposed on a single coefficient. But, when it comes to a joint hypothesis, restrictions are imposed on multiple regression coefficients.

For testing joint hypothesis, we can use the STATA software. Created in 1985 by StataCorp, STATA is a general-purpose interactive statistical software package. The terms 'Statistics' and 'Data' have been truncated and joined to provide the name STATA.

In the current times, STATA has been built for several platforms like Macintosh, Windows and UNIX (Whereas UNIX is UNiplexed Information Computing System (UNICS), later known as UNIX). It comes with various and varied capabilities.

It is capable of performing regression, data management, statistical analysis as also creating graphics and simulations.

It even allows for custom programming and has a built in system for the dissemination of programs written by users. The second capability is a significant contributor to its continuous growth.

NOTES

It provides for an extremely updated coverage of statistical methodology. It is also extremely flexible when it comes to user defined module implementation. This has added to its being preferred software for research as well as statistical analysis, for examining data patterns as well as for graphical data visualisation.

In this unit, you will study about the linear restriction, testing joint hypothesis, problems and applications using STATA.

10.1 OBJECTIVES

After going through this unit, you will be able to:

- Describe the linear restriction
- Analyse the testing joint hypothesis
- Understand the problems and applications using STATA

10.2 INTRODUCTION TO LINEAR RESTRICTIONS

We consider tests on general linear restrictions on regression coefficients. We examine some specific tests of linear restrictions, including the following facts

1. ***F*-test for Regression Fit:** A joint F-test on regression coefficients for all the explanatory variables. The hypotheses are:

   H_0: $\beta_1 = 0$; $\beta_2 = 0$; $\beta_3 = 0$; ....; $\beta_k = 0$

   H_A: At least one β_j ‘– 0

   This is a test with k linear restrictions, all of which are exclusion restrictions.

2. ***F*-Test for a Subset of Regression Coefficients:** A joint F-test on exclusion restrictions for a subset of regression coefficients. For example, suppose we jointly consider variables x_2 and x_3. The hypotheses are:

   H_0: $\beta_2 = 0$; $\beta_3 = 0$

   H_A: At least one of β_2 ‘– 0 and β_3 ‘– 0

   The p-value for the F-test for regression fit is included in the summary regression output in R.

For both of the above tests, the null hypothesis restricts the regression model and the alternative hypothesis is the general unrestricted regression. Both of these tests involve two regressions where one is a restricted test is nested in the unrestricted test. Whenever one has nested regression models and wishes to compare the explanatory power of the unrestricted model versus the restricted model, the following F-test applies following equation:

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)}$$

Where q is equal to the number of restrictions, SSR_r is the residual sum of squares from the restricted regression model, SSR_u is the residual sum of squares from the unrestricted model, n is the sample size, and k is the number of explanatory variables in the unrestricted model.

10.2.1 Simple Linear Regression Model (SLRM)

In a Simple Linear Regression Model (SLRM), hypothesis testing is based on the assumption that an estimator of the regression model is equal to a true value of the parameter when the analysis moves to multiple regression, hypothesis testing varies in the sense that, whether it is advisable to test different slope parameters with separate hypothesis or the researcher should test the same hypothesis for all the slope parameters in the given regression model.

Econometricians are concerned with estimating the regression model with linear restrictions and testing of linear restriction. For example, in the case of Cob Douglas production function the hypothesis of constant returns to scale is similar to the restriction that the sum of the coefficient of the inputs is unity. The simplest type of linear restriction in the regression model is the one which specifies that one or more regression coefficients are equal to zero or addition of other coefficient. In multiple regression models like:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{u}_i$$

The restriction can be specified as

$$\beta_1 + \beta_2 = 0,\ \beta_2 = \beta_3,\ \beta_3 = \beta_4$$

For testing the significance of the estimated partial regression coefficients individually, that is, using the distinct hypothesis that each true population partial regression coefficient was zero.

Assume the below hypothesis:

$$H_0: \beta_2 = \beta_3 = 0 \qquad (10.\ 1)$$

This H_0 null hypothesis is stated as a joint hypothesis specifying that β_2 and β_3 are jointly or simultaneously equal to zero. Testing of these hypothesis is called a test of the overall significance of the estimated regression line, leading to the discussion, if, Y is linearly related to both X_2 and X_3. The joint hypothesis in Equation 10.1 cannot be tested by testing the significance of $\hat{\beta}_2$ and $\hat{\beta}_3$ individually. The reason being that in testing the individual significance of an estimated partial regression coefficient an implicit assumption implies that each test of significance is based on an independent sample. Hence, while testing the significance of $\hat{\beta}_2$ using the hypothesis $\beta_2 = 0$, assumption was made that the testing was based on a different sample from the one used in testing the significance of $\hat{\beta}_3$ with the null hypothesis stating that $\beta_3 = 0$.

While testing the joint hypothesis of Equation 10.1, if the same sample data is considered, the assumption underlying the test procedure that in the given sample Cov ($\hat{\beta}_2$ and $\hat{\beta}_3$) may not be zero is violated.

If we use the same sample data to establish a confidence interval for β_2 and β_3, taking a confidence coefficient of 95%, we cannot declare that both β_2 and β_3 lie in their individual confidence intervals with a probability of linear restriction are

NOTES

$(1 - \alpha)(1 - \alpha) = (0.95)\ (0.95)$

Differently, although the statements

$\Pr [\hat{\beta}_2 - t_{\alpha/2}\ \text{se}\ (\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2}\ \text{se}\ (\hat{\beta}_2)] = 1 - \alpha$

$\Pr [\hat{\beta}_3 - t_{\alpha/2}\ \text{se}\ (\hat{\beta}_3) \leq \beta_3 \leq \hat{\beta}_3 + t_{\alpha/2}\ \text{se}\ (\hat{\beta}_3)] = 1 - \alpha$

Are individually true, *it is wrong to state that* the probability that the intervals

$[\hat{\beta}_2 \pm t_{\alpha/2}\ \text{se}\ (\hat{\beta}_2),\ \hat{\beta}_3 \pm t_{\alpha/2}\ \text{se}\ (\hat{\beta}_3)]$

Simultaneously include β_2 and β_3 is $(1 - \alpha)^2$, since the intervals may not be independent while deriving estimators from the same data set.

Hence, joint hypothesis testing is different from individual hypothesis. While testing several hypotheses jointly, information of one hypothesis will affect another.

10.2.2 Joint Test for the ANalysis Of VAriance (ANOVA): The *F*-Test

The **ANalysis Of VAriance (ANOVA)** Method for the Estimated Multiple Regression for the *F*-test. The usual *t*-test is not useful to test the joint hypothesis that the true partial β- coefficients are zero simultaneously. However, this joint hypothesis can be tested by using the ANalysis Of VAriance (ANOVA) methodology.

The identity

$$\Sigma y_i^2 = \hat{\beta}_2 \Sigma y_i x_{2i} + \hat{\beta}_3 \Sigma y_i x_{3i} + \Sigma \hat{u}_{2i} \qquad (10.2)$$

TSS = ESS + RSS

Total Sum of Square has, *n*-1 degrees of freedom, Residual Sum of Squares has *n*- 3 degrees of freedom and Explained Sum of Squares has 2 degrees of freedom since there are two slope coefficients $\hat{\beta}_2$ and $\hat{\beta}_3$. Therefore, following the ANOVA procedure Table 10.1 is prepared.

Table 10.1 *ANOVA Table for the Three-Variable Regression*

Source of variation	SS	df	MSS
Due to regression (ESS)	$\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}$	2	$\frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{2}$
Due to residual (RSS)	$\sum \hat{u}_i^2$	$n - 3$	$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-3}$
Total	$\sum y_i^2$	$n - 1$	

With the assumption of normal distribution for u_i and the null hypothesis $\beta_2 = \beta_3 = 0$, the variable can be mentioned as the *F*-distribution with 2 and *n*-3 degrees of freedom.

$$F = (\hat{}\beta_2\Sigma y_i x_{2i} + \hat{}\beta_3\Sigma y_i x_{3i})/2 \quad (10.3)$$

$\Sigma\text{–}\hat{}u_{2i}$ / (n-3)

= ESS/Degree of Freedom

NOTES

RSS/ Degree of Freedom

If, the null hypothesis is rejected, that is, X_2 and X_3 certainly effect Y and the Explained sum of squares will be comparatively larger than the Residual sum of squares at the given degrees of freedom. Therefore, the F-value in Equation 10.3 explains a test of the null hypothesis that the true slope, i.e., the β-coefficients are simultaneously zero. If the F-value computed from above mentioned Equation 10.3 exceeds the critical F-value from the F-table at the level of significance, null hypothesis is rejected; otherwise not. Instead, if the p-value of the observed F is sufficiently small, null hypothesis can be rejected. Aforementioned F-testing technique can be generalised as follows.

Decision Rule for Testing the Overall Significance of a Multiple Regression: The *F*-Test

For the k-variable regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i$$

10.2.3 Testing the Hypothesis

A **'Statistical Hypothesis'** is a hypothesis that is testable on the basis of observed data modelled as the realised values taken by a collection of random variables. A set of data is modelled as being realised values of a collection of random variables having a joint probability distribution in some set of possible joint distributions. The hypothesis being tested is exactly that set of possible probability distributions. A statistical hypothesis test is a method of statistical inference. An alternative hypothesis is proposed for the probability distribution of the data, either explicitly or only informally. The comparison of the two models is deemed statistically significant if, according to a threshold probability—the significance level—the data would be unlikely to occur if the null hypothesis were true. A hypothesis test specifies which outcomes of a study may lead to a rejection of the null hypothesis at a pre-specified level of significance, while using a pre-chosen measure of deviation from that hypothesis (the test statistic, or goodness-of-fit measure). The pre-chosen level of significance is the maximal allowed 'False Positive Rate'. One wants to control the risk of incorrectly rejecting a true null hypothesis. The process of distinguishing between the null hypothesis and the alternative hypothesis is aided by considering two types of errors. A Type I error occurs when a true null hypothesis is rejected. A Type II error occurs when a false null hypothesis is not rejected.

H_0: $\beta_2 = \beta_3 = \cdots = \beta_k = 0$ (i.e., all slope coefficients are simultaneously zero)

H_1: Not all slope coefficients are simultaneously zero

NOTES

Calculate

F = ESS/ Degree of Freedom /RSS/ Degree of Freedom

= ESS/(k– 1)/RSS/(n– k) (10.4)

If $F > F_{\alpha}(k-1, n-k)$, reject H_0; otherwise you do not reject it,

Where $F_{\alpha}(k-1, n-k)$ is the *critical* F- value at the α level of significance and (k– 1) and (n– k) are degrees of freedom for numerator and denominator . Otherwise, if the p-value of F calculated from Equation 10.4 is sufficiently low, reject H_0. It must be kept into mind that degree of freedom will depend upon number of β- coefficients. It is worth mentioning that most regression packages like STATA routinely calculate the F-value along with the usual regression output, standard errors and p-values.

10.2.4 Testing the Equality of Two Regression Coefficients

Consider in the following multiple regression

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \qquad (10.5)$$

Testing the hypotheses

$$H_0: \beta_3 = \beta_4 \text{ or } (\beta_3 - \beta_4) = 0 \qquad (10.6)$$

$$H_1: \beta_3 \neq \beta_4 \text{ or } (\beta_3 - \beta_4) \neq 0$$

That is, the two slope coefficients β_3 and β_4 are equal.

This kind of null hypothesis is of applied importance. For example, assume Equation 10.5 is the demand function for a good where Y = amount of a good demanded, X_2 = price of the good, X_3 = income of the consumer, and X_4 = wealth of the consumer.

The null hypothesis must be in this case that the income and wealth coefficients are the indifferent. Or, if Y_i and the X's are logarithmic function, the null hypothesis in Equation 10.6 infers that the income and wealth elasticities of consumption are the same.

Under the classical assumptions, such null hypothesis can be tested as follows:

$$t = \frac{(\hat{\beta}_3 - \hat{\beta}_4) - (\beta_3 - \beta_4)}{\text{se}\,(\hat{\beta}_3 - \hat{\beta}_4)} \qquad (10.7)$$

Follows the t-distribution with (n– 4) Degree of Freedom (DF) because Equation 10.5 is a four-variable model or, more generally, with (n– k) Degree of Freedom where k is the total number of estimators, including the constant term. The se $(\hat{\beta}_3 - \hat{\beta}_4)$ is calculated using the below mentioned formula,

$$\text{se}\,(\hat{\beta}_3 - \hat{\beta}_4) = \sqrt{\text{var}\,(\hat{\beta}_3) + \text{var}\,(\hat{\beta}_4) - 2\,\text{cov}\,(\hat{\beta}_3, \hat{\beta}_4)} \qquad (10.8)$$

Substituting the expression for the se ($\hat{\beta}_3 - \hat{\beta}_4$) into Equation 10.7, the test statistic will be,

NOTES

$$t = \frac{\hat{\beta}_3 - \hat{\beta}_4}{\sqrt{\text{var}(\hat{\beta}_3) + \text{var}(\hat{\beta}_4) - 2\,\text{cov}(\hat{\beta}_3, \hat{\beta}_4)}} \quad (10.9)$$

Testing Procedure

1. Estimate $\hat{\beta}_3$ and $\hat{\beta}_4$ with the help of STATA
2. Calculate the variances and covariance of the estimated parameters using software. Variance can help in estimating standard error in the denominator of Equation 10.9.
3. Find the *t*-ratio from Equation 10.9.
4. Considering the null hypothesis $(\beta_3 - \beta_4) = 0$ if the *t* variable computed exceeds the critical *t*-value at the desired level of significance and available degrees of freedom, reject the null Hypothesis; else do not reject it. Or if the *p*-value of the *t*-statistic is significantly low, one can reject the null hypothesis.
5. Use personal judgement for level of significance at 1, 5, 10 percent and interpretation of the results.

Example 10.1: Assume a Cubic Cost Function as estimated below

$\hat{Y}_i = 141.7667 + 63.4777X_i - 12.9615X_{2i} + 0.9396X_{3i}$

se = (6.3753) (4.7786) (0.9857) (0.0591)

cov ($\hat{\beta}_3$, $\hat{\beta}_4$) = –0.0576; $R^2 = 0.9983$

Where *Y* is total cost and *X* is output,

Suppose the test hypothesis is that the coefficients of the X_2 and X_3 terms in the

Cubic cost function are the same, that is, $\beta_3 = \beta_4$ or $(\beta_3 - \beta_4) = 0$. Following the test procedure

$t = \hat{\beta}_3 - \hat{\beta}_4 / (\text{var}(\hat{\beta}_3) + \text{var}(\hat{\beta}_4) - 2\text{cov}(\hat{\beta}_3, \hat{\beta}_4))^{1/2}$

$= -12.9615 - 0.9396 / ((0.9867)2 + (0.0591)2 - 2(-0.0576))^{1/2}$

$= -13.9011/1.0442$

$= -13.3130$

With 6 degrees of freedom the observed *t*-value exceeds the critical *t*-value even at 0.2 percent level of significance (two-tail test); the *p*-value is extremely small, 0.000006. Hence reject the null hypothesis that the coefficients of X_2 and X_3 in the cubic cost function are equal.

10.2.5 Testing Linear Equality Restrictions: Restricted Least Squares

NOTES

Depending on economic theory at times in a regression model estimated parameter satisfy some linear equality restrictions. Considering the example of the Cobb–Douglas production function:

$$Y_i = \beta_1 X^{\beta 2}_{2i}\, X^{\beta 3}_{3i}\, e^{ui} \qquad (10.10(a))$$

Where Y = output, X_2 = labor input, and X_3 = capital input. Rewriting in log Equation (10.10a)

$$\ln Y_i = \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + ui \qquad (10.10(b))$$

Where $\beta_0 = \ln \beta_1$.

Now if there are constant returns to scale as per economic theory

$$\beta_2 + \beta_3 = 1$$

This is an example of a linear equality restriction.

To validate the restrictions there are two approaches can be taken like a *t*-test and *F*-test.

The *t*-Test Approach

The best and easy technique is to estimate Equation 10.10 usually not taking into account the restriction explicitly. This is referred to as the **unrestricted** or **unconstrained regression.**

Estimating β_2 and β_3 (Using Ordinary least Square technique), a test of the hypothesis or restriction) can be performed by the *t*-test as mentioned below

$$t = \frac{(\hat{\beta}_2 + \hat{\beta}_3) - (\beta_2 + \beta_3)}{\text{se}\,(\hat{\beta}_2 + \hat{\beta}_3)}$$

$$= \frac{(\hat{\beta}_2 + \hat{\beta}_3) - 1}{\sqrt{\text{var}\,(\hat{\beta}_2) + \text{var}\,(\hat{\beta}_3) + 2\,\text{cov}\,(\hat{\beta}_2 \hat{\beta}_3)}} \qquad (10.11)$$

Where $(\beta_2 + \beta_3) = 1$ under the null hypothesis

If the *t*-value computed from Equation10.11 exceeds the critical *t*-value at the given level of significance, reject the null hypothesis of constant returns to scale, otherwise not.

***F*-Test Method**

The *t*-test in the previous section is a kind of post analysis because its effort was to examine whether the linear restriction is satisfied after estimating the 'Unrestricted Regression'.

A more direct approach is to include the restriction into the estimating procedure from the beginning. In the Cobb-Douglas, for example

$$\beta_2 + \beta_3 = 1$$

$$\beta_2 = 1 - \beta_3 \text{ or}$$

$$\beta_3 = 1 - \beta_2$$

Using either one of these equalities, eliminate one of the β-coefficients in Equation 10.10(b) and estimate the equation. Hence, using $\beta_2 = 1 - \beta_3$ write the

Cobb–Douglas production function as

$$\ln Y_i = \beta_0 + (1 - \beta_3) \ln X_{2i} + \beta_3 \ln X_{3i} + u_i$$

$$= \beta_0 + \ln X_{2i} + \beta_3 (\ln X_{3i} - \ln X_{2i}) + u_i$$

Or

$$(\ln Y_i - \ln X_{2i}) = \beta_0 + \beta_3 (\ln X_{3i} - \ln X_{2i}) + u_i \quad (10.12)$$

Or

$$\ln (Y_i/X_{2i}) = \beta_0 + \beta_3 \ln (X_{3i}/X_{2i}) + u_i \quad (10.13)$$

Where (Yi/X_{2i}) = output/labor ratio and (X_{3i}/X_{2i}) = capital labor ratio

Notice how Equation 10.10(b) is transformed. Estimating β_2 from Equations 10.12 or 10.13 β_3 can be easily estimated from the relation $\beta_3 = 1 - \beta_2$

This technique will ensure that the sum of the estimated coefficients of the two inputs will equal 1. The estimation by Equations 10.12 and 10.13 is called as **Restricted Least Squares (RLS).** This procedure can be generalized to models containing any number of explanatory variables and more than one linear equality restriction.

Comparison of Unrestricted and Restricted Least Squares

The best way to compare unrestricted and restricted least square regression is by applying F- test

$\Sigma \hat{u}^2_{UR}$ = RSS of the unrestricted regression

$\Sigma \hat{u}^2_{R}$ = RSS of the restricted regression

When

m= Number of linear restrictions (1 in the Cobb Douglas case above)

k = Number of parameters in the unrestricted regression

n = Number of observations

Then,

$$F = \frac{(\text{RSS}_\text{R} - \text{RSS}_\text{UR})/m}{\text{RSS}_\text{UR}/(n-k)} = \frac{(\sum \hat{u}^2_\text{R} - \sum \hat{u}^2_\text{UR})/m}{\sum \hat{u}^2_\text{UR}/(n-k)}$$

Follows the F-distribution with m, $(n - k)$ degree of freedom.

(*Note:* UR and R stand for unrestricted and restricted, respectively.)

The F-test above can also be expressed in terms of R^2 as follows:

$$F = \frac{(R^2_\text{UR} - R^2_\text{R})/m}{(1 - R^2_\text{UR})/(n-k)}$$

Where R^2_{UR} and R^2_R are, respectively, the R^2 values obtained from the unrestricted and restricted regressions.

NOTES

10.3 STATA

STATA is a general-purpose statistical software package developed by STATACorp for data manipulation, visualisation, statistics, and automated reporting. It is used by researchers in many fields, including economics, sociology, political science, biomedicine, and epidemiology.

Stata was initially developed by Computing Resource Centre in California and the first version was released in 1985. In 1993, the company moved to College Station, TX and was renamed Stata Corporation, now known as StataCorp. A major release in 2003 included a new graphics system and dialog boxes for all commands. Since then, a new version has been released once every two years. The current version is Stata 17, released in April 2021.

10.3.1 Example of Restricted and Unrestricted Regression with STATA

The manual way to establish joint significance for a model is to run an 'Unrestricted Regression' – one which comprises all the variables identified and further run a 'Restricted Regression' – one where variables with small *t*- scores are eliminated. The standard we use to determine whether the variables are jointly significant is whether the increase in variation in model due to residual sum of squares is depicted by STATA software which grows significantly as measured by the *F*-distribution.

Consider an example where one wants to understand the impact of experience, gender, years of education and the interaction of the two on hourly wages. To normalise the data we generate log of hourly wages. The data available in excel is transported to STATA. Follow the steps given below to run the data in STATA

Set-up

```
. gen exper2=exper^2
. gen hwage= wklywage/ wklyhrs
. gen lhwage=log( hwage)
. gen fem=(sex==2)
. gen fexper=fem*exper2
. gen fexper2=fem*exper2
```

The 'Unrestricted Regression'

Command = reg lhwage exper exper2 yrseduc fem fexper fexper2

NOTES

Results

```
reg lhwage exper exper2 yrseduc fem fexper fexper2

      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  6,   993) =   98.74
       Model |  115.86953      6  19.3115884           Prob > F      =  0.0000
    Residual |  194.215486   993  .195584578           R-squared     =  0.3737
-------------+------------------------------           Adj R-squared =  0.3699
       Total |  310.085016   999  .310395411           Root MSE      =  .44225

------------------------------------------------------------------------------
      lhwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0555183   .0042959    12.92   0.000     .0470882    .0639483
      exper2 |  -.0008826   .0000882   -10.01   0.000    -.0010557   -.0007095
     yrseduc |   .0773751   .0053197    14.54   0.000     .0669358    .0878143
         fem |  -.0329736   .0617353    -0.53   0.593    -.1541203    .0881731
      fexper |  -.0263973   .0065542    -4.03   0.000    -.0392589   -.0135357
     fexper2 |   .0003684   .0001361     2.71   0.007     .0001012    .0006355
       _cons |   .5645578   .0801279     7.05   0.000     .4073183    .7217972
------------------------------------------------------------------------------
```

The 'Restricted Regression' – Testing whether fexper and fexper2 are jointly significant

Command = reg lhwage exper exper2 yrseduc fem

Results

```
reg lhwage exper exper2 yrseduc fem

      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  4,   995) =  136.63
       Model |  109.934536     4  27.4836341           Prob > F      =  0.0000
    Residual |  200.150479   995  .201156261           R-squared     =  0.3545
-------------+------------------------------           Adj R-squared =  0.3519
       Total |  310.085016   999  .310395411           Root MSE      =   .4485

------------------------------------------------------------------------------
      lhwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0435066   .0032881    13.23   0.000     .0370541    .0499591
      exper2 |  -.0007107   .0000687   -10.34   0.000    -.0008456   -.0005758
     yrseduc |   .0789745   .0053867    14.66   0.000     .0684038    .0895452
         fem |  -.3188769   .0285882   -11.15   0.000    -.3749769   -.2627768
       _cons |   .6709668   .0785873     8.54   0.000     .5167509    .8251828
------------------------------------------------------------------------------
```

For the regression equation:

lhwage= $\beta_0 + \beta_1$ exper $+ \beta_2$ exper^2 $+ \beta_3$ yrseduc $+ \beta_4$ fem $+ \beta_5$ fexper+ β_6 fexper$^2 + \varepsilon$

The null hypothesis is:

$H_O: \beta_5 = \beta_6 = 0$

Compute the F-statistic:

Fq, N-k = (RSSR – RSSUR) / q / RSSUR / (N-k)

Where:

RSS_{UR} = Residual sum of squares, unrestricted

RSS_R = Residual sum of squares, restricted

q = Number of restrictions (here, the number of variables set equal to zero)

N = Population size

k = Number of variables in the regression, including the constant

NOTES

Here, we get:

F = (200.150479 - 194.215) / 2 / 194.215 / (1000 – 7) = = 15.1725

F-distribution is used to test hypotheses regarding more than one regression parameter. This distribution is appropriate in this case, because the ratio of two Chi-Square distributed variables is distributed as *F* with the two degrees of freedom corresponding to the degrees of freedom in the numerator and denominator of the ratio, respectively.

The computed test statistic is the ratio of sums of squares, and is therefore the ratio of two Chi-Square variables. The *F*-distribution generates always positive values. In above case the computed *F*-statistic will always be positive because the residual sum of squares cannot be made smaller by restricting the regression. The *F*-test is testing whether the residual sum of squares is made 'Significantly' larger by imposing the restriction or whether the regression equation fits roughly equally well with or without the restrictions. Hence the null hypothesis not reject.

To test hypothesis, compare the computed value of *F* to the critical value $F_{q, N-k}$ for a particular at 5% level of significance where 'q' is the number of restrictions imposed on the regression equation. To take a decision on whether the computed value exceeds the critical value and the reject of null hypothesis.

Check Your Progress

1. Which software was created by StataCorp in 1985?
2. Which two words have been truncated and joined to create the bane STATA?
3. Provide 2 examples of platforms for which STATA has been built.
4. List some of the capabilities of STATA.
5. Provide one significant reason for STATA's continuous growth.
6. In a simple linear regression model, what is the assumption for hypothesis testing?
7. In multiple regression, how does hypothesis testing vary?
8. Who is concerned with estimating the regression model with linear restrictions and testing of linear restriction?
9. What is the simplest type of linear restriction in the regression model?
10. The usual *t*-test is not useful to test which joint hypothesis?
11. Which hypothesis can be tested by using the ANOVA methodology?
12. What is the full form of ANOVA?
13. What is 'Unrestricted' regression:
14. What is 'Restricted' regression?
15. Between 'Unrestricted Regression' and 'Restricted Regression', which should be run first?

10.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

NOTES

1. The STATA software was created in 1985 by STATACorp.
2. The terms 'Statistics' and 'Data' have been truncated and joined to provide the name STATA.
3. Macintosh, Windows and UNIX are the examples of STATA plateform.
4. STATA is capable of performing regression, data management, statistical analysis, creating graphics and simulations, custom programming and a built in system for the dissemination of programs written by users.
5. STATA's built in system for the dissemination of programs written by users is a significant contributor to its continuous growth.
6. In a simple linear regression model, hypothesis testing is based on the assumption that an estimator of the regression model is equal to a true value of the parameter.
7. In multiple regression, hypothesis testing is based on whether different slope parameters should be tested with separate hypotheses or the same hypothesis in the given regression model.
8. An econometricians is concerned with estimating the regression model with linear restrictions and testing of linear restriction.
9. The simplest type of linear restriction in the regression model is the one which specifies that one or more regression coefficients are equal to zero or addition of other coefficient.
10. The usual t-test is not useful to test the joint hypothesis that the true partial β coefficients are zero simultaneously.
11. However, this joint hypothesis can be tested by using the ANOVA methodology.
12. The full form of ANOVA is ANalysis Of VAriance.
13. An 'Unrestricted Regression' is one which comprises all the variables identified.
14. A 'Restricted Regression' is one where variables with small t scores are eliminated.
15. The 'Unrestricted Regression' should be fun first, followed by the 'Restricted Regression'.

10.5 SUMMARY

- In a simple linear regression model, hypothesis testing is based on the assumption that an estimator of the regression model is equal to a true value of the parameter.

NOTES

- In multiple regression, hypothesis testing is base on whether it is advisable to test different slope parameters with separate hypotheses or the same hypothesis for all the slope parameters in the given regression model.
- Econometricians are concerned with estimating the regression model with linear restrictions and testing of linear restriction.
- In Cobb-Douglas production function, the hypothesis of constant returns to scale is similar to the restriction that the sum of the coefficient of the inputs is unity.
- The simplest type of linear restriction in the regression model specifies that one or more regression coefficients are equal to zero or addition of other coefficient.
- In multiple regression models like $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{u}_i$, to test the significance of the estimated partial regression coefficients individually, the restriction can be specified as $\beta_{1+} \beta_2 = 0$, $\beta_2 = \beta_{3,}$ $\beta_3 = \beta_4$
- In the hypothesis, H_0 null hypothesis is stated as a joint hypothesis specifying that β_2 and β_3 are jointly or simultaneously equal to zero. Testing of these hypotheses is referred to as a test of the overall significance of the estimated regression line, leading to H_0: $\beta_2 = \beta_3 = 0$,the discussion, if, Y is linearly related to both X_2 and X_3, the aforementioned joint hypothesis cannot be tested by testing the significance of $\hat{\beta}_2$ and $\hat{\beta}_3$ individually.
- When testing the individual significance of an estimated partial regression coefficient, an implicit assumption implies that each test of significance is based on an independent sample.
- While testing the joint hypothesis H_0: $\beta_2 = \beta_3 = 0$, if the same sample data is considered, the assumption underlying the test procedure, that in the given sample Cov ($\hat{\beta}_2$ and $\hat{\beta}_3$) may not be zero is violated.
- If the same sample data is used to establish a confidence interval for β2 and β3, taking a confidence coefficient of 95%, it cannot be declared that both β2 and β3 lie in their individual confidence intervals with a probability of $(1 - \alpha)\ (1 - \alpha) = (0.95)\ (0.95)$.
- While testing several hypotheses jointly, information of one hypothesis affects another.
- The Analysis of Variance (ANOVA) method for joint test for estimated multiple regression uses the F-test.
- In the identity, $\Sigma y^2_i = \hat{\beta}_2 \Sigma y_i x_{2i} + \hat{\beta}_3 \Sigma y_i x_{3i} + \Sigma \hat{u}_{2i}$, TSS = ESS + RSS, Total Sum of Square has n-1 degrees of freedom, Residual Sum of Squares has n-3 degrees of freedom and Explained Sum of Squares has 2 degrees of freedom since there are two slope coefficients: $\hat{\beta}_2$ and $\hat{\beta}_3$.
- The F-testing technique can be generalised as decision rule for testing the overall significance of a multiple regression.

- For the *k*-variable regression model, if $F > F_\alpha(k-1, n-k)$, reject hypothesis that all slope coefficients are simultaneously zero; otherwise you do not reject it.
- When testing the equality of two regression coefficients, in multiple regression, testing that the two slope coefficients, β3 and β4, are equal is a null hypothesis which is of applied importance.
- While testing linear equality restrictions using Restricted Least Squares (RLS), depending on economic theory, at times in a regression model, estimated parameters satisfy some linear equality restrictions, such as in the Cobb-Douglas production function.
- There are two approaches that can be taken to validate the restrictions – *t*-test approach and *F*-test approach.
- The best way to compare unrestricted and restricted least square regression is by applying *F*-test
- The manual way to establish joint significance for a model is to run an 'Unrestricted' regression followed by a 'Restricted' regression.
- The standard for determining if the variables are jointly significant is whether the increase in variation in model due to residual sum of squares is depicted by STATA software which grows significantly as measured by the *F*-distribution.
- The *F*-distribution is used to test hypotheses regarding more than one regression parameter.

10.6 KEY WORDS

- **Linear regression model**: Linear regression follows the linear mathematical model to determining the value of a dependent variable from value of a given independent variable.
- **Hypothesis**: An idea which is suggested to be a possible explanation for something but has not been proven true as yet.
- ***F*-test**: An f-test is any statistical test where under the null hypothesis the test statistic has an *F*-distribution.
- ***t*-test**: A t-test is a type of inferential statistic which is applied to ascertain whether there exists a significant difference between the means of two groups, some of whose features may be related.
- **Regression model**: A regression model helps investigate the relationship between two or more variables and estimate one variable based on the others.
- **Multiple regression**: Otherwise referred to as Multiple Linear Regression (MLR), this statistical technique employs various explanatory variables to predict the outcome of a response variable.

NOTES

- **Cobb-Douglas production function:** It is a functional form of the production function, widely used to represent the technological relationship between the amounts of two or more inputs (particularly physical capital and labor) and the amount of output that can be produced by those inputs.
- **Regression coefficient**: It is an estimate of the unknown population parameters and describe the relationship between a predictor variable and the response.
- **Residual Sum of Squares**: This measures the variation in the error between the observed data and modelled values.
- **Explained Sum of Squares**: This measures how much variation there is in the modelled values.
- **Null hypothesis**: This hypothesis proposes that there is no difference between certain characteristics of a population or of a data-generating process.
- **Degrees of freedom**: This refers to the maximum number of logically independent values in a data sample.
- **STATA**: STATA software can be employed for testing joint hypothesis. STATA is a general-purpose interactive statistical software package. The terms 'statistics' and 'data' have been truncated and joined to provide the name STATA.

10.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Explain the following hypothesis: H_0: $\beta_2 = \beta_3 = 0$
2. In linear regression, what does linear represent?
3. In a linear model, when the estimate of the parameter vector is written as ˆβ=Σiwiyi, what is {wi}?
4. When conducting individual *t*–tests, on how many coefficients is a restriction is imposed?
5. When working with joint hypothesis, on how many regression coefficients are restrictions imposed?

Long-Answer Questions

1. Explain the hypothesis H_0: $\beta_2 = \beta_3 = 0$.
2. In the following identity, list the degrees of freedom for TSS, ESS and RSS.

   $\Sigma y^2_i = \hat{}\beta_2 \Sigma y_i x_{2i} + \hat{}\beta_3 \Sigma y_i x_{3i} + \Sigma \hat{}u_{2i}$

   TSS = ESS + RSS

3. What is the testing procedure for testing the equality of two regression coefficients?
4. Explain how the STATA software is of use to an econometrician.
5. Describe the why the usual *t*-test is not useful to test the joint hypothesis that the true partial β coefficients are zero simultaneously?
6. Explain in detail *F*-test with an example.

NOTES

10.8 FURTHER READINGS

Johnston, J. and John DiNARDO. 1997. *Econometric Methods*, Fourth Edition. New Delhi: Tata McGraw-Hill.

Koutsoyiannis, A.1977. *Theory of Econometrics,* Second Edition. London: The Macmillan Press Ltd.

Özdemir, Durmuº. 2016. *Applied Statistics for Economics and Business,* Second Edition. Izmir (Turkey): Springer.

Maddala, G. S. 1992. *Introduction to Econometrics,* Second Edition. New York: Macmillan Publishing Company.

Pindyck, R. S and D. L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts,* Fourth Edition. New York: McGraw Hill.

Goldberger, A. S. 1998. *Introductory Econometrics***.** Cambridge: Harvard University Press.

Levine, David M., Timothy C. Krehbiei, Mark L. Berenson and P. K. Viswanathan. 2009. *Business Statistics*, Fifth Edition. New Delhi: Pearson Education.

Webster, Allen L. 1998. *Applied Statistics for Business and Economics,* Third Edition. New Delhi: Tata McGraw-Hill.

UNIT 11 TESTING OF HYPOTHESIS

NOTES

Structure

11.0 INTRODUCTION

A hypothesis is an assumption that is tested to find its logical or empirical consequence. A hypothesis should be clear and accurate. Null and alternative hypotheses enable the user to verify the testability of an assumption. It is possible to determine whether hypothesis is appropriate or not with the help of hypothesis tests, such as parametric and non-parametric tests.

Statistical hypothesis testing requires several assumptions. These assumptions include considerations of the level of measurement of the variable, the method of sampling, the shape of the population distribution, and the sample size.

Specification and Hypothesis Tests. We can use the estimated equation to perform hypothesis tests on the coefficients of the model. Specification tests are devised for a number of model specifications in econometrics. Local power is calculated for small departures from the null hypothesis. An instrumental variable test as well as tests for a time series cross section model and the simultaneous equation model are presented. An empirical model provides evidence that unobserved individual factors are present which are not orthogonal to the included right-hand-side variable in a common econometric specification of an individual wage equation.

Hypothesis and prediction are both a type of guess. That's why many people get the two confused. However, the hypothesis is an educated, testable guess in science. A prediction uses observable phenomena to make a future projection.

The application of hypothesis testing in quality management should be promoted. Both parametric test (t-test and z-test) and nonparametric test (sign test and Wilcoxon rank-sum test) are appropriate for use in a manufacturing environment.

In this unit, you will study about the assumption and specification of hypothesis testing, testing of hypothesis prediction, application of hypothesis testing.

11.1 OBJECTIVES

NOTES

After going through this unit, you will be able to:

- Understand the assumption and specification of hypothesis testing
- Analyse the testing of hypothesis prediction
- Discuss about the application of hypothesis testing

11.2 ASSUMPTIONS AND SPECIFICATION OF HYPOTHESIS TESTING

A statistical hypothesis is a hypothesis that is testable on the basis of observed data modelled as the realised values taken by a collection of random variables. A set of data is modelled as being realised values of a collection of random variables having a joint probability distribution in some set of possible joint distributions. The hypothesis being tested is exactly that set of possible probability distributions. A statistical hypothesis test is a method of statistical inference. An alternative hypothesis is proposed for the probability distribution of the data, either explicitly or only informally. The comparison of the two models is deemed statistically significant if, according to a threshold probability the significance level and the data would be unlikely to occur if the null hypothesis were true. A hypothesis test specifies which outcomes of a study may lead to a rejection of the null hypothesis at a pre-specified level of significance, while using a pre-chosen measure of deviation from that hypothesis (the test statistic, or goodness-of-fit measure). The pre-chosen level of significance is the maximal allowed 'False Positive Rate'. One wants to control the risk of incorrectly rejecting a true null hypothesis. The process of distinguishing between the null hypothesis and the alternative hypothesis is aided by considering two types of errors. A Type I error occurs when a true null hypothesis is rejected. A Type II error occurs when a false null hypothesis is not rejected. Hypothesis tests based on statistical significance are another way of expressing confidence intervals (more precisely, confidence sets). In other words, every hypothesis test based on significance can be obtained via a confidence interval, and every confidence interval can be obtained via a hypothesis test based on significance.

A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. It can contain either a suggested explanation for a phenomenon or a proposal having deductive reasoning to suggest a possible interrelation between multiple phenomena. A deductive reasoning can be defined as a type of reasoning that can be derived from previously known facts. The term hypothesis is derived from the ancient Greek term, *hyposthenia*,

which means 'To Put Under' or 'To Suppose'. There are several characteristics of hypothesis, which are:

NOTES

1. **Clear and accurate:** Hypothesis should be clear and accurate so as to draw a consistent conclusion.
2. **Statement of Relationship between Variables:** If a hypothesis is relational, it should state the relationship between different variables.
3. **Testability:** A hypothesis should be open to testing, so that other deductions can be made from it and can be confirmed or disproved by observation. The researcher should do some prior study to make the hypothesis testable.
4. **Specific With Limited Scope:** A hypothesis, which is specific, with limited scope, is easily testable than a hypothesis with limitless scope. Therefore, a researcher should pay more time to do research on such kind of hypothesis.
5. **Simplicity:** A hypothesis should be stated in the most simple and clear terms to make it understandable.
6. **Consistency:** A hypothesis should be reliable and consistent with established and known facts.
7. **Time–Limit:** A hypothesis should be capable of being tested within a reasonable time. In other words, the excellence of a hypothesis is judged by the time taken to collect the data needed for the test.
8. **Empirical Reference:** A hypothesis should explain or support all the sufficient facts needed to understand what the problem is about.

   Hypothesis testing, an analyst tests a statistical sample, with the goal of providing evidence on the plausibility of the null hypothesis. Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analysed. All analysts use a random population sample to test two different hypotheses: the **null hypothesis** and the **alternative hypothesis**. The null hypothesis is usually a hypothesis of equality between population parameters; e.g., a null hypothesis may state that the population mean return is equal to zero. The alternative hypothesis is effectively the opposite of a null hypothesis (e.g., the population mean return is not equal to zero). Thus, they are mutually exclusive, and only one can be true. However, one of the two hypotheses will always be true.

4 Steps of Hypothesis Testing

All hypotheses are tested using a four-step process:

1. The first step is for the analyst to state the two hypotheses so that only one can be right.
2. The next step is to formulate an analysis plan, which outlines how the data will be evaluated.
3. The third step is to carry out the plan and physically analyse the sample data.

4. The fourth and final step is to analyse the results and either reject the null hypothesis, or state that the null hypothesis is plausible, given the data.

Testability of Hypotheses

NOTES

Hypothesis is usually considered as the principal instrument in research. The basic concepts regarding the testability of a hypothesis are as follows:

Null Hypothesis and Alternative Hypothesis

In the context of statistical analysis, the following concepts or assumptions are taken into consideration:

- **Null Hypothesis**: While comparing two different methods in terms of their superiority, wherein the assumption is that both the methods are equally good is called null hypothesis. It is also known as statistical hypothesis and is symbolised as H_0.
- **Alternate Hypothesis**: While comparing two different methods, regarding their superiority, wherein, stating a particular method to be good or bad as compared to the other one is called alternate hypothesis. It is symbolised as H_a. Let us assume that you want to compare a hypothesised mean (μ) with population mean (μ_{H0}), then you will mark null hypothesis as:

  H_0: $= \mu_{H_0} = 100$, where hypothesised mean is equal to population mean.

  Table 11.1 shows the meaning of alternate hypothesis in different situations.

Alternate Hypothesis	Meaning
H_a: $\mu \neq \mu_{H0}$	Population mean is not equal to 100; it could be more or less than 100.
H_a: $\mu > \mu_{H0}$	Population mean is greater than 100.
H_a: $\mu < \mu_{H0}$	Population mean is less than 100.

11.2.1 Assumptions of Hypothesis Testing

Statistical hypothesis testing requires several assumptions. These assumptions include considerations of the level of measurement of the variable, the method of sampling, the shape of the population distribution, and the sample size. The specific assumptions may vary, depending on the test or the conditions of testing. However, without exception, all statistical tests assume random sampling. Tests of hypotheses about means also assume interval-ratio level of measurement and require that the population under consideration be normally distributed or that the sample size be larger than 50. Based on our data, we can test the hypothesis that the average price of gas in India is higher than the average national price of gas. The test we are considering meets these conditions:

1. The sample of Indian gas stations was randomly selected.
2. The variable price per gallon is measured at the interval-ratio level.

3. We cannot assume that the population is normally distributed. However, because our sample size is sufficiently large (N > 50), we know, based on the central limit theorem, that the sampling distribution of the mean will be approximately normal.

Different hypothesis tests make different assumptions about the distribution of the random variable being sampled in the data. These assumptions must be considered when choosing a test and when interpreting the results.

For example, the *z*-test (*z*-test) and the *t*-test (*t*-test) both assume that the data are independently sampled from a normal distribution. Statistics and Machine Learning Toolbox™ functions are available for testing this assumption, such as chi2gof, jbtest, lillietest, and normplot.

Both the *z*-test and the *t*-test are relatively robust with respect to departures from this assumption, so long as the sample size n is large enough. Both tests compute a sample mean x, which, by the Central Limit Theorem (CLT), has an approximately normal sampling distribution with mean equal to the population mean μ, regardless of the population distribution being sampled.

The difference between the *z*-test and the *t*-test is in the assumption of the standard deviation σ of the underlying normal distribution. A *z*-test assumes that σ is known; a *t*-test does not. As a result, a *t*-test must compute an estimate *s* of the standard deviation from the sample.

Test statistics for the *z*-test and the *t*-test are, respectively,

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Under the null hypothesis that the population is distributed with mean μ, the *z*-statistic has a standard normal distribution, $N(0,1)$. Under the same null hypothesis, the *t*-statistic has Student's *t*-distribution with $n - 1$ degrees of freedom. For small sample sizes, Student's *t*-distribution is flatter and wider than $N(0, 1)$, compensating for the decreased confidence in the estimate *s*. As sample size increases, however, Student's *t*-distribution approaches the standard normal distribution, and the two tests become essentially equivalent. Knowing the distribution of the test statistic under the null hypothesis allows for accurate calculation of *p*-values. Interpreting *p*-values in the context of the test assumptions allows for critical analysis of test results.

11.2.2 Specification of Hypothesis Testing

In econometrics, specification tests have been constructed to verify the validity of one specification at a time. These tests will 'Confirm' the validity (or invalidity) of a general model requiring the estimates of the restricted model only.

Using the result that under the null hypothesis of no misspecification an asymptotically efficient estimator must have zero asymptotic covariance with its difference from a consistent but asymptotically inefficient estimator, specification tests are devised for a number of model specifications in econometrics. Local power is calculated for small departures from the null hypothesis. An instrumental variable test as well as tests for a time series cross section model and the simultaneous equation model are presented. An empirical model provides evidence that unobserved individual factors are present which are not orthogonal to the included right-hand-side variable in a common econometric specification of an individual wage equation.

Hypothesis tests consist of following types of specifications:

- Specification of the null hypothesis, H_0.
- Specification of the alternative hypothesis, H_1.
- Specification of the test statistic and its distribution under the null hypothesis.
- Selection of the significance level α in order to determine the rejection region.
- Calculation of the test statistic from the data sample.
- Conclusions, which are based on the test statistic and the rejection region.

11.3 TESTING OF HYPOTHESIS AND PREDICTION

A claim or hypothesis about the values or population parameters is known as the Null Hypothesis and is written as H_0. In the case of the above discussed situation, our assumption that a butler is innocent would form the null hypothesis and would be stated as follows:

$$H_0 = \text{The butler is innocent}$$

This hypothesis is then tested with the available evidence and the decision is made whether to accept this hypothesis or reject it. If this hypothesis is rejected, then we accept the alternate hypothesis which is that the butler is not innocent. This alternate hypothesis is denoted as H_1 and is stated as:

$$H_1 = \text{The butler is not innocent}$$

The process involves testing of the null hypothesis. If the null hypothesis is rejected, then the alternate hypothesis is accepted. It should be noted that the acceptance of the alternate hypothesis does not mean that it is correct. It simply means that there is not enough evidence to be reasonably sure that the null hypothesis is acceptable.

As already explained, there are two types of errors that can be used in making decisions regarding accepting or rejecting the null hypothesis. The first type of error, known as Type I error is used when the null hypothesis is rejected even if it is true. The second type of error, known as Type II error is used when a null hypothesis is accepted even if it was not true and should have been rejected.

NOTES

In statistical hypothesis testing and decision-making about the values of population parameters as defined by the sample statistics, the null hypothesis asserts that there is no true difference between the sample statistics and the corresponding population parameter under consideration and if indeed there is any visible difference, it is considered to be due to natural fluctuations in sampling.

To conclude we say that,

- *Null hypothesis* H_0– An assertion about the population parameter that is being tested by the sample results.
- *Alternate hypothesis* H_1 – A claim about the population parameter that is accepted when the null hypothesis is rejected.
- *Type I error* — An error made in rejecting the null hypothesis, when in fact it is true.
- *Type II error* – An error made in accepting the null hypothesis, when in fact it is false.

Type I error is denoted by α (Alpha) and is expressed as a probability of rejecting a true hypothesis. It is also known as the level of significance. $1 - \alpha$ expresses the level of confidence. For example, $\alpha = 0.05$ means that the confidence level is 95% or 0.95.

Type II error is denoted by β (Beta) and is expressed as the probability of accepting a false hypothesis. It is desirable to have the β value as low as possible for its value reflects the power of the test being performed and a low β value indicates that the test of significance is powerful and reliable.

Procedure For Hypothesis Testing

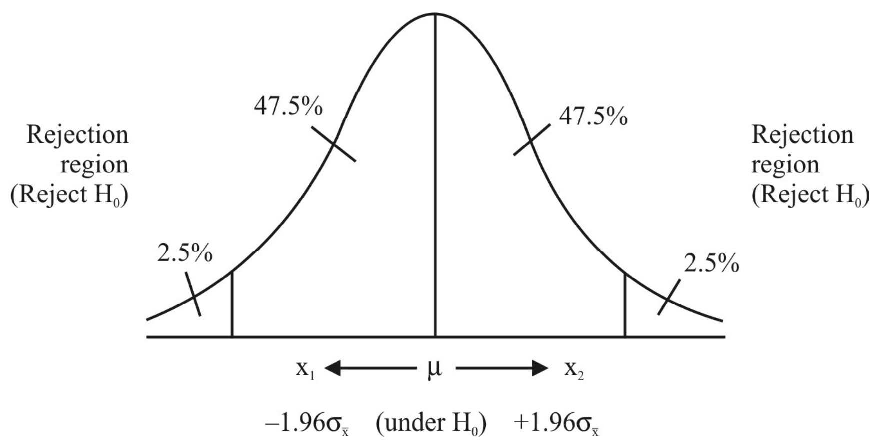
The general procedure for hypothesis testing consists of the following steps:

1. ***State the null hypothesis as well as the alternate hypothesis.*** This means stating the assumed value of the population parameter which is to be tested. For example, suppose that we want to test the hypothesis that the average IQ of our college students is 130. Then this would become our null hypothesis and the alternate hypothesis would be that this average IQ is not 130. These statements are expressed as follows:

$$H_0 : \mu = 130$$
$$H_1 : \mu \neq 130$$

2. ***Establish a level of significance prior to sampling.*** The level of significance signifies the probability of committing Type I error α and is generally taken as equal to 0.05, which really means that after the hypothesis has been tested and a decision is made, we will still be making an error in rejecting the null hypothesis when in fact it is true, 5% of the time. Sometimes the value α is established as 0.01, but it is at the discretion of the investigator to select its value, depending upon the sensitivity of the study.

3. ***Determine a suitable test statistic.*** This means the choice of appropriate probability distribution to use with the particular available information under consideration. The normal distribution using the Z score table or the *t*-distribution is most often used.
4. ***Define the rejection (critical) regions.*** The critical region will be established on the basis of the choice of the value of the level of significance α. For example, if we select the value of $\alpha = 0.05$, and we use the standard normal distribution as our test statistic for testing the population parameter μ, then as we have discussed before, the difference between the assumption of null hypothesis, assumed value of this population parameter and the value obtained by the analysis of sample results is not expected to be more than $\pm 1.96\sigma_{\bar{X}}$ at $\alpha = 0.05$. This relationship can be shown by the following figure.



In the above figure, if the sample $\overline{X}$ statistic falls within $1.96\sigma_{\bar{X}}$ of the assumed value of μ under the assumption of null hypothesis H_0, then we accept the null hypothesis as being correct at 95% confidence level (or 0.05 level of significance). The difference between $\bar{X}$ and μ which may be any value between X_1 and μ or X_2 and μ is considered to be accidental or due to chance element and is not considered significant enough or real enough to reject null hypothesis, so that for all practical purposes the value of $\bar{X}$ is considered equal to μ even though $\bar{X}$ can have any value between X_1 and X_2 as shown above. However, if the value of $\bar{X}$ falls beyond X_2 on the upper side or beyond X_1 on the lower side, then this difference between the values of $\bar{X}$ and μ would be considered significant and it will lead to rejection of null hypothesis. Since 5% of the time, this difference between the values of $\bar{X}$ and μ would be significant with 2.5% of the time $\bar{X}$ being too far above μ (beyond X_2) and 2.5% of the time being too far below μ (below X_1), the area of rejection will be on both sides of the mean extending into the tail sections of the curve. This area of rejection is known as the *critical region.*

NOTES

5. ***Data collection and sample analysis.*** This involves the actual collection and computation of the sample data. A sample of the pre-established size *n* is collected and the estimate of the population parameter is calculated. This estimate is the value of the test statistic. For example, if we are testing a hypothesis about the value of population mean μ, then the test statistic would be the sample mean $\bar{X}$. Then we test this statistic to check whether it falls in the critical region or in the acceptance region. For example, if we want to test for the average IQ of the college students to be 130, then in that case we have to see that our population mean μ must be tested. We take a random sample of a given size *n* and calculate its mean $\bar{X}$ and then test it to see if the value of this $\bar{X}$ falls in the area of acceptance or in the area of rejection at a given level of significance.
6. ***Making the decision.*** Before the statistical decision is made, a decision rule must be established. Such decision rule will form the basis on which the null hypothesis will be accepted or rejected. This decision rule is really a formal statement of the obvious purpose of the test. For example, this rule could be stated as follows,

   Accept the null hypothesis if the value of sample statistic $\bar{X}$ *falls within the area of acceptance, otherwise reject the null hypothesis.*

   Based upon this established decision rule, a decision can be made whether to accept or reject the null hypothesis.

Committing Errors: Type I and Type II

- **Types of Errors**: There are two types of errors in statistical hypothesis, which are as follows:
  - o **Type I Error**: In this type of error, you may reject a null hypothesis when it is true. It means rejection of a hypothesis, which should have been accepted. It is denoted by α (alpha), and is also known alpha error.
  - o **Type II Error**: In this type of error, you are supposed to accept a null hypothesis when it is not true. It means accepting a hypothesis, which should have been rejected. It is denoted by β (beta), and is also known as beta error.

Type I error can be controlled by fixing it at a lower level, for example, If you fix it at 2%, then the maximum probability to commit Type I error is 0.02. But reducing Type I error, has a disadvantage when the sample size is fixed as it increases the chances of Type II error. In other words, it can be said that both types of errors cannot be reduced simultaneously. The only solution of this problem is to set an appropriate level by considering the costs and penalties attached to them or to strike a proper balance between both types of errors.

In a hypothesis test, a type I error occurs when the null hypothesis is rejected when it is in fact true; that is, H_0 is wrongly rejected. For example, in a clinical trial

of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; that is H_0: there is no difference between the two drugs on average. A type I error would occur if we concluded that the two drugs produced different effects when in fact there was no difference between them.

In a hypothesis test, a type II error occurs when the null hypothesis H_0, is not rejected when it is in fact false. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; that is H_0: there is no difference between the two drugs on average. A type II error would occur if it were concluded that the two drugs produced the same effect, that is, there is no difference between the two drugs on average, when in fact, they produced different ones.

In how many ways can we commit errors?

We reject a hypothesis when it may be true. This is Type I Error.

We accept a hypothesis when it may be false. This is Type II Error.

The other true situations are desirable:

We accept a hypothesis when it is true. We reject a hypothesis when it is false.

	Accept H_0	**Reject H_0**
H_0 True	Accept True H_0 Desirable	Reject True H_0 Type I Error
H_1 False	Accept False H_0 Type II Error	Reject False H_0 Desirable

The level of significance implies the probability of type I error. A five per cent level implies that the probability of committing a type I error is 0.05. A one per cent level implies 0.01 probability of committing type I error.

Lowering the significance level and hence the probability of type I error is good but unfortunately it would lead to the undesirable situation of committing type II error.

To sum up:

- **Type I Error**: Rejecting H_0 when H_0 is true.
- **Type II Error**: Accepting H_0 when H_0 is false.

Note: The probability of making a Type I error is the level of significance of a statistical test. It is denoted by α.

Where, α = Prob. (Rejecting H_0 / H_0 true)

$1 - \alpha$ = Prob. (Accepting H_0 / H_0 true)

The probability of making a Type II error is denoted by β.

Where, β = Prob. (Accepting H_0 / H_0 false)

$1 - \beta$ = Prob. (Rejecting H_0 / H_0 false) = Prob. (The test correctly rejects H_0 when H_0 is false)

NOTES

$1 - \beta$ is called the power of the test. It depends on the level of significance α, sample size n and the parameter value.

Null and Alternative Hypothesis

Hypothesis is usually considered as the principal instrument in research. The basic concepts regarding the testability of a hypothesis are as follows:

Null Hypothesis and Alternative Hypothesis

In the context of statistical analysis, while comparing any two methods, the following concepts or assumptions are taken into consideration:

- **Null Hypothesis**: While comparing two different methods in terms of their superiority, wherein the assumption is that both the methods are equally good is called null hypothesis. It is also known as statistical hypothesis and is symbolized as H_0.
- **Alternate Hypothesis**: While comparing two different methods, regarding their superiority, wherein, stating a particular method to be good or bad as compared to the other one is called alternate hypothesis. It is symbolized as H_1.

Comparison of Null Hypothesis with Alternate Hypothesis

Following are the points of comparison between null hypothesis and alternate hypothesis:

- Null hypothesis is always specific while alternate hypothesis gives an approximate value.
- The rejection of null hypothesis involves great risk, which is not in the case of alternate hypothesis.

Null hypothesis is more frequently used in statistics than alternate hypothesis because it is specific and is not based on probabilities.

The hypothesis to be tested is called the Null Hypothesis and is denoted by H_0. This is to be tested against other possible states of nature called alternative hypothesis. The alternative is usually denoted by $H_{1.}$

The null hypothesis implies that there is no difference between the statistic and the population parameter. To test whether there is no difference between the sample mean X and the population m, we write the null hypothesis.

$$H_0: \; \overline{X} = \mu$$

The alternative hypothesis would be,

$$H_1: \neq \mu$$

This means > μ or < μ. This is called a two-tailed hypothesis.

The alternative hypothesis H_1: > μ is right tailed.

The alternative hypothesis H_1: < μ is left tailed.

These are one sided or one-tailed alternatives.

Note 1: The alternative hypothesis H_1 implies all such values of the parameter, which are not specified by the null hypothesis H_0.

Note 2: Testing a statistical hypothesis is a rule, which leads to a decision to accept or reject a hypothesis.

A one-tailed test requires rejection of the null hypothesis when the sample statistic is greater than the population value or less than the population value at a certain level of significance.

1. We may want to test if the sample mean exceeds the population mean m. Then the null hypothesis is,

   H_0: > μ

2. In the other case the null hypothesis could be,

   H_0: < μ

Each of these two situations leads to a one-tailed test and has to be dealt with in the same manner as the two-tailed test. Here the critical rejection is on one side only, right for > μ and left for < μ. Both the Figures 11.3 and 11.4 here show a five per cent level of test of significance.

For example, a minister in a certain government has an average life of 11 months without being involved in a scam. A new party claims to provide ministers with an average life of more than 11 months without scam. We would like to test if, on the average, the new ministers last longer than 11 months. We may write the null hypothesis H_0: = 11 and alternative hypothesis H_1: > 11.

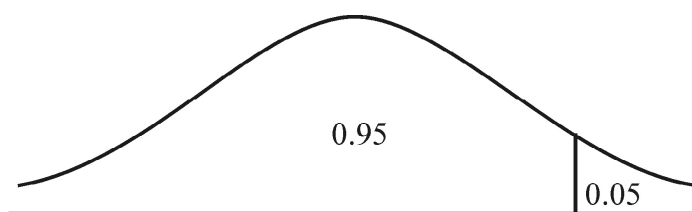


Fig. 11.3 H_0: $\overline{X}$ > μ

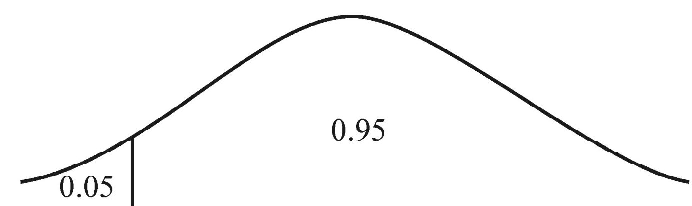


Fig. 11.4 H_0: < μ

11.4 APPLICATIONS

NOTES

- Application of hypothesis testing will allow manufacturers to better understand quality data and provide guidance to production control.
- Hypothesis testing substantiates that the use of only a descriptive statistic, such as arithmetic mean, sum and range, fails to provide a panoramic view of product or service quality.
- Real world applications of hypothesis testing include:
  - Testing whether more men than women suffer from nightmares
  - Establishing authorship of documents
  - Evaluating the effect of the full moon on behaviour
  - Determining the range at which a bat can detect an insect by echo
  - Deciding whether hospital carpeting results in more infections
  - Selecting the best means to stop smoking
  - Checking whether bumper stickers reflect car owner behaviour
  - Testing the claims of handwriting analysts
- It helps to provide link to the underlying theory and specific research question. It helps in data analysis and measure the validity and reliability of the research. It provides a basis or evidence to prove the validity of the research.
- Testing of hypothesis, also known as sample-testing, is a common feature with almost every social and management research. We draw conclusion on population (characteristics) based on available sample information, following certain statistical principles.
- Applications of hypothesis testing for environmental science presents the theory and application of hypothesis testing in environmental science, allowing researchers to carry out suitable tests for decision-making on a variety of issues. The tests are presented in simplified form without relying on complex mathematical proofs to allow researchers to easily locate the most appropriate test and apply it to real-world situations. Each example is accompanied by a case study showing the application of the method to realistic data.
- Hypothesis testing can be used in business applications to help validate an assumption being made about data relationships.
- Hypothesis testing is a mathematical tool for confirming a financial or business claim or idea.
- Hypothesis testing is useful for investors trying to decide what to invest in and whether the instrument is likely to provide a satisfactory return.

NOTES

- Hypothesis testing is also used by manufacturing and quality engineers where you have to sample a value from a process or a production line to try to figure out if the process is at the nominal value or drifting.
- Hypothesis testing is the process used to evaluate the strength of evidence from the sample and provides a framework for making determinations related to the population, i.e., it provides a method for understanding how reliably one can extrapolate observed findings in a sample under study to the larger population from.
- Hypothesis testing is an integral and most important component of research methodology, in all researches, whether in medical sciences, social sciences or any such allied field. It is a guideline in planning, implementation and getting final results thereof, in undertaking any research work.
- For that confession of data, 'Hypothesis Testing' could be used to interpret and draw conclusions about the population using sample data. A hypothesis test helps in making a decision as to which mutually exclusive statement about the population is best supported by sample data.

Check Your Progress

1. Elaborate on the statistical hypothesis.
2. Give the four steps of hypothesis testing.
3. What is null hypothesis?
4. Explain about the assumption of hypothesis testing.
5. Give the types of specification consist of hypothesis testing.
6. Name the two types of errors in statistical hypothesis.
7. Why is null hypothesis more frequently used in statistics?
8. Interpret the application of hypothesis testing in environmental science.
9. Give the medical application of hypothesis testing.

11.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. A statistical hypothesis is a hypothesis that is testable on the basis of observed data modelled as the realised values taken by a collection of random variables.
2. 4 Steps of Hypothesis Testing

   All hypotheses are tested using a four-step process:

   1. The first step is for the analyst to state the two hypotheses so that only one can be right.

NOTES

2. The next step is to formulate an analysis plan, which outlines how the data will be evaluated.
3. The third step is to carry out the plan and physically analyse the sample data.
4. The fourth and final step is to analyse the results and either reject the null hypothesis, or state that the null hypothesis is plausible, given the data.

3. **Null Hypothesis**: While comparing two different methods in terms of their superiority, wherein the assumption is that both the methods are equally good is called null hypothesis. It is also known as statistical hypothesis and is symbolised as H0.
4. Statistical hypothesis testing requires several assumptions. These assumptions include considerations of the level of measurement of the variable, the method of sampling, the shape of the population distribution, and the sample size. The specific assumptions may vary, depending on the test or the conditions of testing. However, without exception, all statistical tests assume random sampling.
5. Hypothesis tests consist of following types of specifications:
   - Specification of the null hypothesis, H_0.
   - Specification of the alternative hypothesis, H_1.
   - Specification of the test statistic and its distribution under the null hypothesis.
   - Selection of the significance level α in order to determine the rejection region.
   - Calculation of the test statistic from the data sample.
   - Conclusions, which are based on the test statistic and the rejection region.
6. **Types of Errors**: There are two types of errors in statistical hypothesis, which are as follows:
   - o **Type I Error**: In this type of error, you may reject a null hypothesis when it is true. It means rejection of a hypothesis, which should have been accepted. It is denoted by α (alpha), and is also known alpha error.
   - o **Type II Error**: In this type of error, you are supposed to accept a null hypothesis when it is not true. It means accepting a hypothesis, which should have been rejected. It is denoted by β (beta), and is also known as beta error.
7. Null hypothesis is more frequently used in statistics than alternate hypothesis because it is specific and is not based on probabilities.

8. Applications of hypothesis testing for environmental science presents the theory and application of hypothesis testing in environmental science, allowing researchers to carry out suitable tests for decision-making on a variety of issues. The tests are presented in simplified form without relying on complex mathematical proofs to allow researchers to easily locate the most appropriate test and apply it to real-world situations. Each example is accompanied by a case study showing the application of the method to realistic data.
9. Hypothesis testing is an integral and most important component of research methodology, in all researches, whether in medical sciences, social sciences or any such allied field. It is a guideline in planning, implementation and getting final results thereof, in undertaking any research work.

11.6 SUMMARY

- A statistical hypothesis is a hypothesis that is testable on the basis of observed data modelled as the realised values taken by a collection of random variables.
- A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. It can contain either a suggested explanation for a phenomenon or a proposal having deductive reasoning to suggest a possible interrelation between multiple phenomena.
- Hypothesis should be clear and accurate so as to draw a consistent conclusion.
- A hypothesis should be open to testing, so that other deductions can be made from it and can be confirmed or disproved by observation. The researcher should do some prior study to make the hypothesis testable.
- A hypothesis should be capable of being tested within a reasonable time. In other words, the excellence of a hypothesis is judged by the time taken to collect the data needed for the test.
- Hypothesis testing, an analyst tests a statistical sample, with the goal of providing evidence on the plausibility of the null hypothesis. Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analysed.
- The alternative hypothesis is effectively the opposite of a null hypothesis (e.g., the population mean return is not equal to zero). Thus, they are mutually exclusive, and only one can be true.
- While comparing two different methods, regarding their superiority, wherein, stating a particular method to be good or bad as compared to the other one is called alternate hypothesis.
- All statistical tests assume random sampling. Tests of hypotheses about means also assume interval-ratio level of measurement and require that the

NOTES

population under consideration be normally distributed or that the sample size be larger than 50.

- In econometrics, specification tests have been constructed to verify the validity of one specification at a time. These tests will 'Confirm' the validity (or invalidity) of a general model requiring the estimates of the restricted model only.
- In statistical hypothesis testing and decision-making about the values of population parameters as defined by the sample statistics, the null hypothesis asserts that there is
- Type I error is denoted by α (Alpha) and is expressed as a probability of rejecting a true hypothesis. It is also known as the level of significance. $1 - \alpha$ expresses the level of confidence.
- Type II error is denoted by β (Beta) and is expressed as the probability of accepting a false hypothesis. It is desirable to have the β value as low as possible for its value reflects the power of the test being performed and a low β value indicates that the test of significance is powerful and reliable.
- In a hypothesis test, a type I error occurs when the null hypothesis is rejected when it is in fact true; that is, H_0 is wrongly rejected.
- In a hypothesis test, a type II error occurs when the null hypothesis H_0, is not rejected when it is in fact false.
- The hypothesis to be tested is called the Null Hypothesis and is denoted by H_0. This is to be tested against other possible states of nature called alternative hypothesis. The alternative is usually denoted by H_1.
- Testing a statistical hypothesis is a rule, which leads to a decision to accept or reject a hypothesis.
- Application of hypothesis testing will allow manufacturers to better understand quality data and provide guidance to production control.
- Testing of hypothesis, also known as sample-testing, is a common feature with almost every social and management research. We draw conclusion on population (characteristics) based on available sample information, following certain statistical principles.
- Hypothesis testing can be used in business applications to help validate an assumption being made about data relationships.
- Hypothesis testing is also used by manufacturing and quality engineers where you have to sample a value from a process or a production line to try to figure out if the process is at the nominal value or drifting.

11.7 KEY WORDS

- **Statistical hypothesis:** A statistical hypothesis is a hypothesis that is testable on the basis of observed data modelled as the realised values taken by a collection of random variables.
- **Null hypothesis:** While comparing two different methods in terms of their superiority, wherein the assumption is that both the methods are equally good is called null hypothesis. It is also known as statistical hypothesis and is symbolised as H_0.
- **Alternate Hypothesis**: While comparing two different methods, regarding their superiority, wherein, stating a particular method to be good or bad as compared to the other one is called alternate hypothesis. It is symbolised as H_a.
- **Specification**: In econometrics, specification tests have been constructed to verify the validity of one specification at a time. These tests will 'Confirm' the validity (or invalidity) of a general model requiring the estimates of the restricted model only.

11.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Define the statistical hypothesis.
2. Give the steps of hypothesis testing.
3. What is alternative hypothesis?
4. Elaborate on the null hypothesis.
5. Interpret the assumptions of hypothesis testing.
6. Explain about the specification of hypothesis testing.
7. Give the procedure of hypothesis testing.
8. Distinguish between null and hypothesis.
9. Give the application of hypothesis testing for medical field.
10. Explain about the hypothesis testing used in business field.

Long-Answer Questions

1. Briefly explain about the hypothesis testing with the appropriate examples.
2. Describe the assumption and specification of hypothesis testing with the help of examples.

3. Explain in detail about the testing of hypothesis and prediction.
4. Analyse the procedure for hypothesis testing giving various types of steps.
5. Discuss in detail about the applications of hypothesis testing in various field.

NOTES

11.9 FURTHER READINGS

Johnston, J. and John DiNARDO. 1997. *Econometric Methods*, Fourth Edition. New Delhi: Tata McGraw-Hill.

Koutsoyiannis, A.1977. *Theory of Econometrics,* Second Edition. London: The Macmillan Press Ltd.

Özdemir, Durmuº. 2016. *Applied Statistics for Economics and Business,* Second Edition. Izmir (Turkey): Springer.

Maddala, G. S. 1992. *Introduction to Econometrics,* Second Edition. New York: Macmillan Publishing Company.

Pindyck, R. S and D. L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts,* Fourth Edition. New York: McGraw Hill.

Goldberger, A. S. 1998. *Introductory Econometrics***.** Cambridge: Harvard University Press.

Levine, David M., Timothy C. Krehbiei, Mark L. Berenson and P. K. Viswanathan. 2009. *Business Statistics*, Fifth Edition. New Delhi: Pearson Education.

Webster, Allen L. 1998. *Applied Statistics for Business and Economics,* Third Edition. New Delhi: Tata McGraw-Hill.

NOTES

BLOCK - IV
ECONOMETRIC METHODS AND SOFTWARE PACKAGES

UNIT 12 ESTIMATION METHODS

Structure

12.0 INTRODUCTION

Methods of estimation in Modern Econometrics provides a comprehensive introduction to a wide range of emerging topics, such as generalised empirical likelihood estimation and alternative asymptotic under drifting parameterisations. Econometric methods, which are statistical estimation techniques and econometric models to which estimation methods are applied. Econometric analysis is used to develop, estimate and evaluate models which relate economic or financial variables.

A variety of methods are used in econometrics to estimate models consisting of a single equation. The oldest and still the most commonly used is the ordinary least squares method used to estimate linear regressions. A variety of methods are available to estimate non-linear models. A particularly important class of non-linear models are those used to estimate relationships where the dependent variable is discrete, truncated or censored. These include logit, probit and Tobit models. Single equation methods may be applied to time-series, cross section or panel data.

The system methods are Three Stages Least Square (3SLS), Iterative Three Stages Least Square (IT3SLS), and Full Information Maximum Likelihood (FIML). These methods use information concerning the endogenous variables in the system and take into account error covariance's across equations and hence are asymptotically efficient in the absence of specification error.

Numerical analysis is the study of algorithms that use numerical approximation (as opposed to symbolic manipulations) for the problems of mathematical analysis (as distinguished from discrete mathematics).

In this unit, you will study about the estimation methods, single equation and systems estimation method and numerical problems.

12.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain about the various types of estimation methods
- Understand the single equation and systems estimation method
- Analyse the numerical problems based on estimation

12.2 ESTIMATION METHODS

Assumptions

Econometric techniques are used to estimate economic models, which ultimately allow you to explain how various factors affect some outcome of interest or to forecast future events. The Ordinary Least Squares (OLS) technique is the most popular method of performing regression analysis and estimating econometric models, because in standard situations (meaning the model satisfies a series of statistical assumptions) it produces optimal (the best possible) results.

The proof that OLS generates the best results is known as the Gauss-Markov theorem, but the proof requires several assumptions. These assumptions, known as the Classical Linear Regression Model (CLRM) assumptions, are the following:

- The model parameters are linear, meaning the regression coefficients don't enter the function being estimated as exponents (although the variables can have exponents).
- The values for the independent variables are derived from a random sample of the population, and they contain variability.
- The explanatory variables do not have perfect collinearity (that is, no independent variable can be expressed as a linear function of any other independent variables).
- The error term has zero conditional mean, meaning that the average error is zero at any specific value of the independent variable(s).
- The model has no heteroscedasticity (meaning the variance of the error is the same regardless of the independent variable's value).

- The model has no autocorrelation (the error term does not exhibit a systematic relationship over time).

Estimation theory is a branch of statistics that deals with estimating the values of parameters based on measured empirical data that has a random component. The estimation parameters describe an underlying physical setting in such a way that their value affects the distribution of the measured data. An estimator attempts to approximate the unknown parameters using the measurements. Read In estimation theory, two approaches are generally considered.

- The probabilistic approach (described in this article) assumes that the measured data is random with probability distribution dependent on the parameters of interest
- The set-membership approach assumes that the measured data vector belongs to a set which depends on the parameter vector.

Commonly used estimators (estimation methods) and topics related to them include:

1. Maximum Likelihood Estimators
2. Bayes Estimators
3. Method of Moment's Estimators
4. Cramér–Rao Bound
5. Least Squares
6. Minimum Mean Squared Error (MMSE), also known as Bayes Least Squared Error (BLSE)
7. Maximum A Posteriori (MAP)

1. **Maximum Likelihood Estimators:** In statistics, Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a probability distribution by maximising a likelihood function, so that under the assumed statistical model the observed data is most probable. The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate. The logic of maximum likelihood is both intuitive and flexible, and, such as the method has become a dominant means of statistical inference. If the likelihood function is differentiable, the derivative test for determining maxima can be applied. In some cases, the first-order conditions of the likelihood function can be solved explicitly; for instance, the ordinary least squares estimator maximizes the likelihood of the linear regression model. Under most circumstances, however, numerical methods will be necessary to find the maximum of the likelihood function. From the vantage point of Bayesian inference, MLE is a special case of maximum a posteriori estimation that assumes a uniform prior distribution of the parameters. In frequentist inference, MLE is a special case of an extremum estimator, with the objective function being the likelihood.

NOTES

2. **Bayes Estimators:** In estimation theory and decision theory, a **Bayes estimator** or a Bayes action is an estimator or decision rule that minimises the posterior expected value of a loss function (i.e., the posterior expected loss). Equivalently, it maximises the posterior expectation of a utility function. An alternative way of formulating an estimator within Bayesian statistics is maximum a posteriori estimation.
3. **Method of Moment's Estimators:** In statistics, the method of moments is a method of estimation of population parameters. It starts by expressing the population moments (i.e., the expected values of powers of the random variable under consideration) as functions of the parameters of interest. Those expressions are then set equal to the sample moments. The number of such equations is the same as the number of parameters to be estimated. Those equations are then solved for the parameters of interest. The solutions are estimates of those parameters. The method of moments was introduced by Pafnuty Chebyshev in 1887 in the proof of the central limit theorem. The idea of matching empirical moments of a distribution to the population moments dates back at least to Pearson.
4. **Cramér–Rao Bound:** In estimation theory and statistics, the Cramér–Rao Bound (CRB) expresses a lower bound on the variance of unbiased estimators of a deterministic (fixed, though unknown) parameter, stating that the variance of any such estimator is at least as high as the inverse of the Fisher information. The result is named in honor of Harald Cramér and C. R. Rao, but has independently also been derived by Maurice Fréchet, Georges Darmois, as well as Alexander Aitken and Harold Silverstone. An unbiased estimator which achieves this lower bound is said to be (fully) efficient. Such a solution achieves the lowest possible mean squared error among all unbiased methods, and is therefore the Minimum Variance Unbiased (MVU) estimator. However, in some cases, no unbiased technique exists which achieves the bound. This may occur either if for any unbiased estimator, there exists another with a strictly smaller variance, or if an MVU estimator exists, but its variance is strictly greater than the inverse of the Fisher information. The Cramér–Rao bound can also be used to bound the variance of biased estimators of given bias. In some cases, a biased approach can result in both a variance and a mean squared error that are below the unbiased Cramér–Rao lower bound.
5. **Least Squares:** The method of least squares is a standard approach in regression analysis to approximate the solution of overdetermined systems (sets of equations in which there are more equations than unknowns) by minimising the sum of the squares of the residuals made in the results of every single equation. The most important application is in data fitting. The best fit in the least-squares sense minimises the sum of squared residuals (a residual being: the difference between an observed value, and the fitted

value provided by a model). When the problem has substantial uncertainties in the independent variable (the x variable), then simple regression and least-squares methods have problems; in such cases, the methodology required for fitting errors-in-variables models may be considered instead of that for least squares.

6. **Minimum Mean Square Error (MMSE):** Estimator In statistics and signal processing, a Minimum Mean Square Error (MMSE) Estimator is an estimation method which minimises the Mean Square Error (MSE), which is a common measure of estimator quality, of the fitted values of a dependent variable. In the Bayesian setting, the term MMSE more specifically refers to estimation with quadratic loss function. In such case, the MMSE estimator is given by the posterior mean of the parameter to be estimated. Since the posterior mean is cumbersome to calculate, the form of the MMSE estimator is usually constrained to be within a certain class of functions. Linear MMSE estimators are a popular choice since they are easy to use, easy to calculate, and very versatile. It has given rise to many popular estimators, such as the Wiener–Kolmogorov filter and Kalman filter.
7. **Maximum A Posteriori (MAP):** In Bayesian statistics, a Maximum A Posteriori Probability (MAP) estimate is an estimate of an unknown quantity, that equals the mode of the posterior distribution. The MAP can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data. It is closely related to the method of Maximum Likelihood (ML) estimation, but employs an augmented optimisation objective which incorporates a prior distribution (that quantifies the additional information available through prior knowledge of a related event) over the quantity one wants to estimate. MAP estimation can therefore be seen as a regularisation of maximum likelihood estimation.

12.3 SINGLE EQUATION AND SYSTEMS ESTIMATION METHOD

The method of statistically drawing an inference on data is called the '**Statistical Inference**'. Thus, the testing of hypothesis and the inference are the most important factors involved. The theory of estimation is a part of statistics that extracts parameters from observations that are corrupted with noise. Estimation is the part of statistics and signal processing that determines the values of parameters through measured and observed empirical data. The process of estimation is carried out in order to measure and diagnose the true value of a function or a particular set of populations. It is done on the basis of observations on the samples, which are a combined piece of the target population or function. Several statistics are used to perform the task of estimation.

12.3.1 Single Equation of Estimation Method

NOTES

A variety of methods are used in econometrics to estimate models consisting of a single equation. The oldest and still the most commonly used is the ordinary least squares method used to estimate linear regressions. A variety of methods are available to estimate non-linear models. A particularly important class of non-linear models are those used to estimate relationships where the dependent variable is discrete, truncated or censored. These include logit, probit and Tobit models.

Single equation methods may be applied to time-series, cross section or panel data. Single equation methods are used in econometrics to estimate models in which a single variable of interest is determined by one or more exogenous explanatory variables.

Traditional Time Series Analysis for a Single Equation

Research using time series data in political science typically has utilised many of the same regression techniques as are employed to analyse cross-sectional data. The vast majority of these traditional time series analyses have considered single-equation models such as the following:

$$Y_t = \beta_0 + \Sigma\beta_{1-k} X_{1-k,t-i} + \varepsilon_t \quad (12.1)$$

Where Y_t is the dependent variable at time t, X_{t-i} are 1 to k independent variables at time $t - i$, β_0 is constant, β_{1-k} are the parameters associated with variables $X_{1 \text{"} k}$, and μ_t is the stochastic error term $\sim N(0, \sigma^2)$.

An econometric model, in the form of a single stochastic equation, is a primary tool in econometrics. The subject of its description consists of a dependent variable Y with y_t observations, where t is the statistical observation's number (t = 1, …, n) and n is the sample size. The dependent variable is economic in character and represents a specific economic category.

Explanatory variables marked as $X_l, \ldots, X_j, \ldots, X_k$, essentially, represent the factors causing variations of the dependent variable Y. Also, some statistical observations are assigned to each dependent variables: x_{tl}, representing the variable $X_l, \ldots, x_{tj}$, representing the variable X_j, … as well as x_{tk} for the variable X_k.

The most general form of a model with a single stochastic equation can be written as follows:

$$y_t = f\left(x_{t1}, \ldots, x_{tj}, \ldots x_{tk}, \eta_t\right), \quad (12.2)$$

With one more variable $\cdot_t$, the random component. This random component gives the model its stochastic character and results from the following:

- The random nature of economic phenomena and processes.
- A conscious and purposeful resignation from complying with less important and statistically insignificant factors.

- Inaccuracies during observation and measurement of economic phenomena and processes.
- A lack of full precision in determining the equation's analytical form.
- Round-ups in the course of numerical calculations.

NOTES

12.3.2 System Estimation Method

There are two fundamental methods of estimation for simultaneous equations: least squares and maximum likelihood. There are two approaches within each of these categories: single equation methods and system estimation. Two Stage Least Square (2SLS), Three Stage Least Square (3SLS), and Iterative Three Stages Least Square (IT3SLS) use the least-squares method; Limited-Information Maximum Likelihood (LIML) and Full Information Maximum Likelihood (FIML) use the maximum likelihood method. 2SLS and LIML are single equation methods, which means that over identifying restrictions in other equations are not taken into account in estimating parameters in a particular equation. System methods are 3SLS, IT3SLS, and FIML. These methods use information concerning the endogenous variables in the system and take into account error covariance's across equations and hence are asymptotically efficient in the absence of specification error.

- *K*-class estimation is a class of estimation methods that include the 2SLS, Ordinary Least Squares (OLS), LIML, and Minimum Expected Loss (MELO) methods as special cases. A *K*-value less than 1 is recommended but not required.
- MELO is a Bayesian *K*-class estimator. It yields estimates that can be expressed as a matrix weighted average of the OLS and 2SLS estimates.
- The Seemingly Unrelated Regressions (SUR) and Iterative Seemingly Unrelated Regressions (ITSUR) methods use information about contemporaneous correlation among error terms across equations in an attempt to improve the efficiency of parameter estimates.

Instrumental Variables and *K*-Class Estimation Methods

Instrumental variable methods involve substituting a predicted variable for the endogenous variable Y when it appears as a regressor. The predicted variables are linear functions of the instrumental variables and the endogenous variable.

The 2SLS method substitutes $\hat{Y}$ for *Y*, which results in consistent estimates. In 2SLS, the instrumental variables are used as regressors to obtain the projected value $\hat{Y}$, which is then substituted for *Y*. Normally, the predetermined variables of the system are used as the instruments. It is possible to use variables other than predetermined variables from your system of equations as instruments; however, the estimation may not be as efficient. For consistent estimates, the instruments must be uncorrelated with the residual and correlated with the endogenous variable.

NOTES

K-class estimators are instrumental variable estimators where the first-stage predicted values take a special form:

$$\mathbf{Y}^* = (1 - k)\mathbf{Y} + k\hat{\mathbf{Y}}$$

For a specified value *k*. The probability limit of *k* must equal 1 for consistent parameter estimates.

The LIML method results in consistent estimates that are exactly equal to 2SLS estimates when an equation is exactly identified. LIML can be viewed as least-variance ratio estimators or as maximum likelihood estimators. LIML involves minimizing the ratio

$$\lambda = (rvar_eq)/(rvar_sys),$$

Where *rvar_eq* is the residual variance associated with regressing the weighted endogenous variables on all predetermined variables appearing in that equation, and *rvar_sys* is the residual variance associated with regressing weighted endogenous variables on all predetermined variables in the system. The *K*-class interpretation of LIML is that $k = \lambda$. Unlike OLS and 2SLS, where *k* is 0 and 1, respectively, *k* is stochastic in the LIML method.

The MELO method computes the minimum expected loss estimator. The MELO method computes estimates that 'Minimise the posterior expectation of generalised quadratic loss functions for structural coefficients of linear structural models' (Judge *et al.* 1985, 635). Other frequently used *K*-class estimators may not have finite moments under some commonly encountered circumstances and hence there can be infinite risk relative to quadratic and other loss functions. MELO estimators have finite second moments and hence finite risk.

One way of comparing *K*-class estimators is to note that when $k = 1$, the correlation between regressor and the residual is completely corrected for. In all other cases, it is only partially corrected for.

SUR and 3SLS Estimation Methods

SUR may improve the efficiency of parameter estimates when there is contemporaneous correlation of errors across equations. In practice, the contemporaneous correlation matrix is estimated using OLS residuals. Under two sets of circumstances, SUR parameter estimates are the same as those produced by OLS: when there is no contemporaneous correlation of errors across equations (the estimate of contemporaneous correlation matrix is diagonal); and when the independent variables are the same across equations.

Theoretically, SUR parameter estimates will always be at least as efficient as OLS in large samples, provided that your equations are correctly specified. However, in small samples the need to estimate the covariance matrix from the

OLS residuals increases the sampling variability of the SUR estimates, and this effect can cause SUR to be less efficient than OLS. If the sample size is small and the across-equation correlations are small, then OLS should be preferred to SUR. The consequences of specification error are also more serious with SUR than with OLS.

The 3SLS method combines the ideas of the 2SLS and SUR methods. Like 2SLS, the 3SLS method uses instead of Y for endogenous regressors, which results in consistent estimates. Like SUR, the 3SLS method takes the cross-equation error correlations into account to improve large sample efficiency. For 3SLS, the 2SLS residuals are used to estimate the cross-equation error covariance matrix.

The SUR and 3SLS methods can be iterated by recomputing the estimate of the cross-equation covariance matrix from the SUR or 3SLS residuals and then computing new SUR or 3SLS estimates based on this updated covariance matrix estimate. Continuing this iteration until convergence produces ITSUR or IT3SLS estimates.

FIML Estimation Method

The FIML estimator is a system generalisation of the LIML estimator. The FIML method involves minimising the determinant of the covariance matrix associated with residuals of the reduced form of the equation system. From a maximum likelihood standpoint, the LIML method involves assuming that the errors are normally distributed and then maximising the likelihood function subject to restrictions on a particular equation. FIML is similar, except that the likelihood function is maximized subject to restrictions on all of the parameters in the model, not just those in the equation being estimated. The FIML method is implemented as an instrumental variable method (Hausman 1975).

Choosing a Method for Simultaneous Equations

A number of factors should be taken into account in choosing an estimation method. Although system methods are asymptotically most efficient in the absence of specification error, system methods are more sensitive to specification error than single equation methods.

In practice, models are never perfectly specified. It is a matter of judgment whether the misspecification is serious enough to warrant avoidance of system methods.

Another factor to consider is sample size. With small samples, 2SLS may be preferred to 3SLS. In general, it is difficult to say much about the small sample properties of K-class estimators because this depends on the regressors used.

LIML and FIML are invariant to the normalization rule imposed but are computationally more expensive than 2SLS or 3SLS.

NOTES

If the reason for contemporaneous correlation among errors across equations is a common omitted variable, it is not necessarily best to apply SUR. SUR parameter estimates are more sensitive to specification error than OLS. OLS may produce better parameter estimates under these circumstances. SUR estimates are also affected by the sampling variation of the error covariance matrix. There is some evidence from Monte Carlo studies that SUR is less efficient than OLS in small samples.

12.4 NUMERICAL PROBLEMS

In statistics, the method of estimating equations is a way of specifying how the parameters of a statistical model should be estimated. This can be thought of as a generalisation of many classical methods the method of moments, least squares, and maximum likelihood as well as some recent methods like M-estimators. The basis of the method is to have, or to find, a set of simultaneous equations involving both the sample data and the unknown model parameters which are to be solved in order to define the estimates of the parameters.Various components of the equations are defined in terms of the set of observed data on which the estimates are to be based. Important examples of estimating equations are the likelihood equations.

In order to evaluate the accuracy of a numerical solution, the backward error analysis algorithm considers the numerical solution as an exact solution of a problem close to the original problem. This distinguishes the backward analysis from the direct error analysis, which estimates the accuracy of the errors in the calculation of numerical solutions. It is known that in the methods of direct error analysis, a large increase in the bounds of estimates is noted, which in most cases greatly exceeds the values of the numerical errors themselves. In the monograph, Stetter argued that 'Direct error analysis is almost never applicable, except for the simplest applications of the discretisation method, and usually you have to rely heavily on the information obtained in the process of computing numerical solutions'. Noticed around 1950 that all difference schemes are incorrect in a certain sense: as h?®0 it necessary to increase the mesh dimension unlimitedly. In fact, for many numerical algorithms, the boundary of the error of the numerical solution (which can be constructed) is growing rapidly, although the error values themselves often turn out to be much smaller. This was the basis for the development of backward analysis problems of the theory of errors: given the error of the function, it is required to determine the errors of its arguments. An unknown- dependent equation with many solutions is constructed. In order to single out one solution to the problem, it is necessary to put an additional condition (or several conditions). Hence, when assessing the accuracy of the obtained numerical solution, this solution is believed to be an exact solution to a problem approximating the original problem. Such a method was developed by Wilkinson J. in problems for the numerical solution of linear algebra problems, and extended to other areas of numerical analysis. Using

NOTES

direct and backward error analysis Voevodin V V effectively computed majorants of rounding errors in the most important methods of linear algebra, and significantly developed the results. If we consider the rounding errors of the results of intermediate calculations as functions that depend on random input data, for direct problems it was proved that the rounding errors asymptotically behave as independent, uniformly distributed random variables (in terms of the number of digits of the representation of numbers). Studies were conducted on the influence of small perturbations of the input data on the solution of many problems of linear algebra, including incorrectly posed ones. Using the backward error analysis, fairly accurate error estimates were obtained for solutions of Systems of Linear Algebraic Equations (SLAE) as well as for many other problems.

The Global Error Estimation Algorithm for Ordinary Differential Equations (ODE) Systems in the Operator Form

Let a system of Ordinary Differential Equations (ODEs) with initial data be solved.

$$F(y) := \begin{pmatrix} -y(t_0) + y_0 \\ -y'(t) + f(t, y(t)) \end{pmatrix}, t_0 \le t \le T \tag{12.3}$$

This system is written in the form of a nonlinear operator equation F(y) = 0, in which all additional conditions are included in the operator that maps some function spaces. The requirements are imposed to ensure the existence and uniqueness of the solution. We apply the numerical method and obtain the approximate solution, which approximates the projection of the exact solution onto the difference grid of the domain of the function argument. The uniform grid constructed for the ODE system is given as $t_n = a + nh;\ n = 0, \ldots.N$ whereas N is a positive integer number. To compare the exact and approximate solutions, one has to project all the solutions into one functional space (discrete or continuous).

Definition 1. The global error of the interpolant of the numerical solution of the ODE system is the difference between the exact solution of the system and the interpolant of the numerical solution. In some cases, the term global error is also used to denote the norm of the difference of these values.

Definition 2. A defect, or residual, is a quantity

$$\delta(t) := \frac{du}{dt}(t) - f(t, u) \tag{12.4}$$

The backward analysis of numerical solutions errors can be performed by applying a defect to compute the bound of global error $\| y(t) - u(t) \|$, where $y(t)$ is an exact solution of the ODE system and $u(t)$ is an interpolant constructed by a numerical solution y_h. Since this means that $\frac{du}{dt} = f(t, u) + \delta(t)$, the defect is equal to the quantity measuring the extent to which the numerical solution does not satisfy the differential Equation 12.3. The concept of a defect (residual) can

also be compared with the difference between the original Equation 12.3 and the equation whose exact solution is a numerical solution. The global error norm $\varepsilon = \|y_h - \Delta y\|$.

NOTES

In the general case, for a sufficiently smooth operator *F,* the error $\varepsilon = O(h^r)$, where *h* is the grid step, *r* is a positive integer characterizing the accuracy order of the numerical solution.

In evaluating the global error, the deferred correction approach is of great importance, consisting in computing a more accurate numerical solution $\overline{y_h}$, obtained by adding the correction term to the numerical solution found earlier. Computation requires the use of some numerical method ϕ. As ϕ, you can use the method by which a numerical solution y_h was obtained, which can be done with cost savings. As a variant of the method ϕ, a more efficient method order p ≥1 can also be used. For example, let the Euler method be chosen for problem.

$$\varphi(y_h)_i := \begin{cases} -y_{h,0} + \alpha, \quad i = 0, \\ -\frac{(y_{h,i} - y_{h,i-1})}{h} + f(t_{i-1}, y_{h,i-1}), \quad i = 1, ..., N \end{cases} \tag{12.5}$$

To implement deferred correction, a local error $\lambda := \phi(\Delta y)$ is used, which characterises the quality of approximation by an exact solution of the problem of solving the difference equation $\phi(y_h)$.

For the Euler method,

$$\lambda_i = \begin{cases} 0, & i = 0 \\ -\frac{hy''(\tau_i)}{2}, & i = 1, ..., N \end{cases}$$

In this formula, τ is some intermediate point on the *i* interval. Obviously, if the magnitude of the local error λ is known with a sufficient degree of accuracy, then it is easy to find the exact solution at the grid nodes Δ*y* by solving the equation $\phi(\Delta y) = \lambda$ The main idea of the deferred correction method is to obtain an estimate of the local error using some operator ψ, an estimator of the local error of the numerical solution y_h. Since we used $\eta = \Delta y + O(h^r)$ to estimate the magnitude of the order h^p, it would seem that at best the estimate would satisfy $\psi(\eta) = \lambda + O(h^{r+p})$. The more precise solution $\overline{\eta}$ is found as a solution of system of equations if it can be expected, for a stable method equations $\phi(\overline{\eta}) = \psi(\eta)$; if $\psi(\eta) = \phi(\Delta y) + O(h^{r+p})$ it can be expected, for a stable method ϕ, $\eta = \Delta y + O(h^p)$ which gives an estimate of the global error $\eta - \overline{\eta}$ accurate to $O(h^{r+p})$. So, deferred correction reduces the global error estimation problem to the local error estimation problem. It can be assumed that the operator Æ must satisfy the relation $\phi(\Delta y) = \lambda + O(h^{r+p})$ However, this is not enough, since ϕ it satisfies this condition precisely and one cannot be sure what $\phi(y_h)$ is a suitable estimate, especially if y_h is found when solving the equation $\phi(y_h) = 0$.

It is necessary to put the second condition on *psi* so that the error $O(h^r)$ when calculating y_h is reduced by $O(h^p)$. Many numerical experiments have shown that this condition is usually satisfied if $\phi(\Delta z) = O(h^p)$ for an arbitrary sufficiently smooth function z.

NOTES

As an example of the estimated function of the local error of the Euler method we consider,

$$\psi(y_h)_i := \begin{cases} 0, & i = 0 \\ -\left(\frac{h}{2}\right) \cdot \left(\frac{f(t_i, y_{h,i}) - f(t_{i-1}, y_{h,i-1})}{h}\right), & i = 1, ..., N. \end{cases}$$

For an arbitrary function z

$$\psi(\Delta z)_n = -\left(\frac{h}{2}\right)\left(f_t(\tau_n', z(\tau_n')) + f_z(\tau_n', z(\tau_n'))z'(\tau_n')\right) O(h)$$

and with $z = y$

$$\psi(\Delta y)_n = -\left(\frac{h}{2}\right) y''(\tau_n') = \lambda_n + O(h^2).$$

As a result, delayed correction has three components: a numerical solution y_h, an effective numerical method ϕ, and an estimated local error function ϕ an improved solution $\overline{y_h}$ is calculated on the basis of the equation $\phi(\overline{y_h}) = \phi(y_h)$ and satisfies the relation $\overline{y_h} = \Delta y + O(h^{r+p})$, provided that for an arbitrary function $y_h = \Delta y + O(h^r)$, $\phi(\Delta y) = \phi(\Delta y) + O(h^{r+p})$ and $\phi(\Delta z) = O(h^p)$.

Regularization in Constructing an Estimate of an Approximate Solution Using a Defect Change

We define the operator equation,

$$A_z = u, z \in Z, u \in U, \qquad (12.6)$$

Where Z, U are some metric spaces. Problem (12.6) is called correctly posed if the following conditions are satisfied: Equation 12.3 is the operator Equation 12.5 is solvable for any right-hand side $u \in U$, is the solution of the operator equation is stable under perturbation of the right-hand side of the Equation 12.6, this means the continuity of the inverse operator A^{-1}; defined over the entire space U, is the solution of operator Equation 12.6 is unique. If at least one of these conditions is not fulfil the problem is called incorrectly posed.

The class of ill-posed problems is very wide; these include the problems of summing Fourier series with coefficients given with errors, some optimal control problems, and some linear algebra problems, and minimising functional. The experience of using direct and backward error analysis methods gives reason to believe that these methods are also incorrect problems; for the backward analysis, the use of regularisation of the problem is justified.

Results of Numerical Experiments

For the oscillation equation (written in the form of a system of two differential equations of the first order) on the interval [0, 10000], the values of the numerical

NOTES

solution were calculated using one-step Runge- Kutta method of the fourth order and the multi-step Adams method of the fourth order Given that the exact solution to the system is known and does not necessitate the use of numerical methods aimed at compensating for complex areas of solutions (for example, singularities or stiffness's), an error estimate (global error) was obtained for the system. We used deferred difference correction and Richardson extrapolation, as well as a reverse error analysis. Estimates calculated using reverse error analysis look like this:

$err\ (10) = 0.456148217\ .10^{-6}$, $err\ (100) = 0.6465776799\ .10^{-5}$,

$err\ (1000) = 0.78572562324.10^{-4}$, $err(10000) = 0.187345329827.\ 10^{-3}$

Let us solve the ODE system,

$$\frac{dy_1(t)}{dt} = y_2(t),\ \frac{dy_2(t)}{dt} = y_3(t),\ \frac{dy_3(t)}{dt} = -y_2(t) - 0.5y_1^2(t) + 1$$

With initial data $y_1(0) = 0, y_2(0) = 0, y_3(0) = 1$ using the one-step Runge- Kutta method of the fourth order and the multi-step Adams method of the fourth order. Error estimates calculated using delayed difference correction or Richardson extrapolation are equal at $t = 1$

$$err_{y1}(1) = .617423026019743194 - .617423026020436306 \approx 10^{-16},$$
$$err_{y2}(1) = .1.29539072722494274 - .1.29539072723032510 \approx 10^{-16},$$
$$err_{y3}(1) = 1.34603247053568098 - 1.34603247053760988 \approx 10^{-12}$$

Error estimates calculated using delayed difference correction or Richardson extrapolation are equal at $t = 7.5$

$$err_{y1}(7.5) \approx 10^{-5}, err_{y1}(7.5) \approx 10^{-4}, err_{y1}(7.5) \approx 10^{-3}$$

Error estimates calculated using the backward error analysis are equal at $t = 7.5$

$$err_{y1}(7.5) \approx 10^{-6}, err_{y1}(7.5) \approx 10^{-6}, err_{y1}(7.5) \approx 10^{-5}$$

Conclusion

The experience of estimating errors in numerical solutions confirm that one of the advantages of the inverse error analysis is that applying it to a well-defined problem with a small inverse error leads to a small direct error (global error for numerical solutions to ODEs). Finding a certain set of finite diameter to which the exact solution belongs (at least for sufficiently small errors) and estimating the distance from the approximate solution to its boundary allows us to reduce the influence of the incorrectness of the inverse error analysis, which manifests itself in the absence of stability. We can characterise the analysis of errors in numerical solutions as an

incorrectly posed problem. The regularisation of the operator of the problem to be solved, which consists in replacing the original system with systems of a simpler form, helps to obtain more accurate error estimates.

NOTES

Check Your Progress

1. Elaborate on the ordinary least square.
2. What do you understand by estimation theory?
3. Explain about the estimation parameter.
4. Interpret the maximum likelihood estimator.
5. What is Bayes estimator?
6. Elaborate on the maximum a posteriori.
7. What is statistical inference?
8. What do you understand by single equation of estimation method?
9. Define the *K*-class estimation.
10. Comprehend the FIML estimation method.
11. Write a short note on estimation equation.

12.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The Ordinary Least Squares (OLS) technique is the most popular method of performing regression analysis and estimating econometric models, because in standard situations (meaning the model satisfies a series of statistical assumptions) it produces optimal (the best possible) results.
2. Estimation theory is a branch of statistics that deals with estimating the values of parameters based on measured empirical data that has a random component.
3. The estimation parameters describe an underlying physical setting in such a way that their value affects the distribution of the measured data. An estimator attempts to approximate the unknown parameters using the measurements.
4. Maximum Likelihood Estimators: In statistics, Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a probability distribution by maximising a likelihood function, so that under the assumed statistical model the observed data is most probable.
5. Bayes Estimators: In estimation theory and decision theory, a Bayes estimator or a Bayes action is an estimator or decision rule that minimises the posterior expected value of a loss function (i.e., the posterior expected loss).

NOTES

6. Maximum A Posteriori (MAP): In Bayesian statistics, a Maximum A Posteriori Probability (MAP) estimate is an estimate of an unknown quantity, that equals the mode of the posterior distribution. The MAP can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data.
7. The method of statistically drawing an inference on data is called the 'Statistical Inference'.
8. Single equation methods are used in econometrics to estimate models in which a single variable of interest is determined by one or more exogenous explanatory variables.
9. *K*-class estimation is a class of estimation methods that include the 2SLS, Ordinary Least Squares (OLS), LIML, and Minimum Expected Loss (MELO) methods as special cases. A *K*-value less than 1 is recommended but not required.
10. The FIML estimator is a system generalisation of the LIML estimator. The FIML method involves minimising the determinant of the covariance matrix associated with residuals of the reduced form of the equation system.
11. In statistics, the method of estimating equations is a way of specifying how the parameters of a statistical model should be estimated. This can be thought of as a generalisation of many classical methods the method of moments, least squares, and maximum likelihood as well as some recent methods like M-estimators.

12.6 SUMMARY

- Econometric techniques are used to estimate economic models, which ultimately allow you to explain how various factors affect some outcome of interest or to forecast future events.
- The proof that OLS generates the best results is known as the Gauss-Markov theorem, but the proof requires several assumptions.
- In statistics, Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a probability distribution by maximising a likelihood function, so that under the assumed statistical model the observed data is most probable.
- In estimation theory and decision theory, a Bayes estimator or a Bayes action is an estimator or decision rule that minimises the posterior expected value of a loss function (i.e., the posterior expected loss).
- In statistics, the method of moments is a method of estimation of population parameters. It starts by expressing the population moments (i.e., the expected values of powers of the random variable under consideration) as functions of the parameters of interest.

NOTES

- In estimation theory and statistics, the Cramér–Rao Bound (CRB) expresses a lower bound on the variance of unbiased estimators of a deterministic (fixed, though unknown) parameter, stating that the variance of any such estimator is at least as high as the inverse of the Fisher information.
- The Cramér–Rao bound can also be used to bound the variance of biased estimators of given bias. In some cases, a biased approach can result in both a variance and a mean squared error that are below the unbiased Cramér–Rao lower bound.
- The method of least squares is a standard approach in regression analysis to approximate the solution of overdetermined systems (sets of equations in which there are more equations than unknowns) by minimising the sum of the squares of the residuals made in the results of every single equation.
- The best fit in the least-squares sense minimises the sum of squared residuals (a residual being: the difference between an observed value, and the fitted value provided by a model).
- Estimator In statistics and signal processing, a Minimum Mean Square Error (MMSE) Estimator is an estimation method which minimises the Mean Square Error (MSE), which is a common measure of estimator quality, of the fitted values of a dependent variable.
- In the Bayesian setting, the term MMSE more specifically refers to estimation with quadratic loss function.
- The theory of estimation is a part of statistics that extracts parameters from observations that are corrupted with noise. Estimation is the part of statistics and signal processing that determines the values of parameters through measured and observed empirical data.
- A variety of methods are used in econometrics to estimate models consisting of a single equation. The oldest and still the most commonly used is the ordinary least squares method used to estimate linear regressions.
- Single equation methods may be applied to time-series, cross section or panel data. Single equation methods are used in econometrics to estimate models in which a single variable of interest is determined by one or more exogenous explanatory variables.
- Research using time series data in political science typically has utilised many of the same regression techniques as are employed to analyse cross-sectional data.
- MELO is a Bayesian K-class estimator. It yields estimates that can be expressed as a matrix weighted average of the OLS and 2SLS estimates.
- The Seemingly Unrelated Regressions (SUR) and Iterative Seemingly Unrelated Regressions (ITSUR) methods use information about contemporaneous correlation among error terms across equations in an attempt to improve the efficiency of parameter estimates.

NOTES

- The FIML estimator is a system generalisation of the LIML estimator. The FIML method involves minimising the determinant of the covariance matrix associated with residuals of the reduced form of the equation system.
- In statistics, the method of estimating equations is a way of specifying how the parameters of a statistical model should be estimated. This can be thought of as a generalisation of many classical methods the method of moments, least squares, and maximum likelihood as well as some recent methods like M-estimators.
- In order to evaluate the accuracy of a numerical solution, the backward error analysis algorithm considers the numerical solution as an exact solution of a problem close to the original problem.
- Using the backward error analysis, fairly accurate error estimates were obtained for solutions of Systems of Linear Algebraic Equations (SLAE) as well as for many other problems.

12.7 KEY WORDS

- **Estimation theory:** Estimation theory is a branch of statistics that deals with estimating the values of parameters based on measured empirical data that has a random component.
- **Bayes estimators:** In estimation theory and decision theory, a Bayes estimator or a Bayes action is an estimator or decision rule that minimises the posterior expected value of a loss function (i.e., the posterior expected loss).
- **Cramér–Rao bound:** In estimation theory and statistics, the Cramér–Rao Bound (CRB) expresses a lower bound on the variance of unbiased estimators of a deterministic (fixed, though unknown) parameter, stating that the variance of any such estimator is at least as high as the inverse of the Fisher information.
- **Statistical inference:** The method of statistically drawing an inference on data is called the 'Statistical Inference'.
- **Single equation method:** Single equation methods may be applied to time-series, cross section or panel data. Single equation methods are used in econometrics to estimate models in which a single variable of interest is determined by one or more exogenous explanatory variables.
- **FIML estimator:** The FIML estimator is a system generalisation of the LIML estimator. The FIML method involves minimising the determinant of the covariance matrix associated with residuals of the reduced form of the equation system.

12.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

NOTES

Short-Answer Questions

1 .What is estimation method?
2. Give the uses of estimation method.
3. State the estimation theory.
4. Elaborate on the Bayes estimator.
5. Define the least square.
6. What do you understand by minimum mean square error?
7. Explain about the single equation of estimation method.
8. Elaborate on the system estimation method.
9. What is SUR and 3SLS estimation method?
10. Explain the estimation equation.
11. Give the definition of global error of estimation.

Long-Answer Questions

1. Briefly explain about the different types of estimation method with the appropriate examples.
2. What is single equation of estimation method? Briefly explain about the time series analysis for a single equation.
3. Describe the system estimation method with the help of examples.
4. Analyse the numerical problems based on estimation method.

12.9 FURTHER READINGS

Johnston, J. and John DiNARDO. 1997. *Econometric Methods*, Fourth Edition. New Delhi: Tata McGraw-Hill.

Koutsoyiannis, A.1977. *Theory of Econometrics,* Second Edition. London: The Macmillan Press Ltd.

Özdemir, Durmuº. 2016. *Applied Statistics for Economics and Business,* Second Edition. Izmir (Turkey): Springer.

Maddala, G. S. 1992. *Introduction to Econometrics,* Second Edition. New York: Macmillan Publishing Company.

NOTES

Pindyck, R. S and D. L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts,* Fourth Edition. New York: McGraw Hill.

Goldberger, A. S. 1998. *Introductory Econometrics***.** Cambridge: Harvard University Press.

Levine, David M., Timothy C. Krehbiei, Mark L. Berenson and P. K. Viswanathan. 2009. *Business Statistics*, Fifth Edition. New Delhi: Pearson Education.

Webster, Allen L. 1998. *Applied Statistics for Business and Economics,* Third Edition. New Delhi: Tata McGraw-Hill.

UNIT 13 DYNAMIC ECONOMETRIC MODELS

NOTES

Structure

13.0 INTRODUCTION

Econometric models are statistical models used in econometrics. An econometric model specifies the statistical relationship that is believed to hold between the various economic quantities pertaining to a particular economic phenomenon. An econometric model can be derived from a deterministic economic model by allowing for uncertainty, or from an economic model which itself is stochastic. However, it is also possible to use econometric models that are not tied to any specific economic theory.

The dynamic econometric models consist of both the lag and the time element in it. Basically, they are of two types: Auto-Regressive (AR) models and Distributed Lag (DL) models. Auto-Regressive (AR) models consist the lagged values of the dependent or endogenous variable. A model called an autoregressive model, if it consist one or more lagged values of the dependent variable among its explanatory variables.

Distributed Lag (DL) models consist the lagged values of the explanatory variables. If the length of the lag is defined, then it is known as 'Finite Distributed Lag Models'. If we don't know the length of the lag or it is infinite, then it is known as 'Infinite Distributed Lag Models'.

NOTES

Dynamic stochastic general equilibrium modelling (abbreviated as DSGE, or DGE, or sometimes SDGE) is a macroeconomic method which is often employed by monetary and fiscal authorities for policy analysis, explaining historical time-series data, as well as future forecasting purposes. DSGE econometric modelling applies general equilibrium theory and microeconomic principles in a tractable manner to postulate economic phenomena, such as economic growth and business cycles, as well as policy effects and market shocks.

In this unit, you will study about the dynamic econometric models, nature and preliminary analysis of economic time series, integration, tests of stationary, unit root test, non-stationary, and the problem of spurious regression.

13.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain the dynamic econometric models
- Define the nature and preliminary analysis of economic time series
- Elaborate on the integration
- Analyse the tests of stationary
- Comprehend the unit root test
- Interpret the non-stationary
- Understand the problem of spurious regression

13.2 ANALYSIS OF ECONOMIC TIME SERIES

The times series data is defined as data values of the variables at equally spaced time interval. The analysis takes into account that the data observations taken over the time period have an internal structure, such as autocorrelation, specific trends or having seasonal variations.

The key objective of time series analysis is to develop an understanding of the underlying phenomenon and structure represented by the observed data. Further, the idea is to fit a model that leads to forecasting, monitoring and feedback, and forward control.

The time series analysis consists of the concepts mentioned below and each will be discussed heuristically for our analysis with examples, wherever applicable.

- Stochastic processes
- Stationarity processes
- Non-stationary processes
- Test of stationarity

- Integrated variables
- Unit root tests
- Spurious regression

NOTES

13.3 STOCHASTIC PROCESSES

A random or stochastic process is an assembly of random variables which are ordered in time. Let *Y* be a random variable, and if it is continuous then denoted it as $Y(t)$, but if it is discrete, denoted it as Y_t. An example of continuous is an electrocardiogram, and some of discrete value are GDP and PDI. Majority of economic data are collected at discrete points in time, hence notation used will be Y_t. If *Y* represents GDP, and having data for 88 quarters $Y_1, Y_2, Y_3, \ldots, Y_{86}, Y_{87}, Y_{88}$, where the subscript 1 denotes the first observation (i.e., assume GDP for the first quarter of 1991) and the subscript 88 denotes the last observation (i.e., GDP for the fourth quarter of 2019). Allow to understand that each of these Ys is a random variable.

GDP is a stochastic variable and to understand the same assume the GDP of $2872.8 billion for 1991–I quarter. According to theory, the GDP value for the first quarter of 1991 could have been any random number, depending on the economic and political environment prevailing in the nation. The figure of 2872.8 is a particular realization of all such possibilities. Hence, GDP is a stochastic process and the actual values observed for the given time are a precise realization of that sample.

Similar to the role of sample data in drawing inferences about a population, in time series econometricians use realization to draw inferences about the underlying stochastic process.

13.4 STATIONARY STOCHASTIC PROCESSES

The stochastic process studied in great detail by time series analysts is the stationary stochastic process. Generally, a stochastic process is considered to be stationary if its mean and variance are constant over time and the value of the covariance between the two time periods depends only on the distance or gap or lag between the two time periods and not the actual time at which the covariance is computed.

In the time series language, such a stochastic process is defined as weakly stationary, or second-order stationary.

To explain second-order stationary, let Y_t be a stochastic time series with the below mentioned properties:

Mean: $E(Y_t) = \mu$ (13.1)

Variance: $var(Y_t) = E(Y_t - \mu)\ 2 = \sigma^2$ (13.2)

Covariance: $\gamma k = E[(Y_t - \mu)(Y_t + k - \mu)]$ (13.3)

NOTES

Where γk, the auto covariance at lag k, is the covariance between the values of Y_t and $Y_t + k$, that is, between two Y values k periods apart.

If $k = 0$, we obtain γ_0, known as the variance of Y (= σ^2);

If $k = 1$, γ_1 is the covariance between two adjacent values of Y.

Assume shifting the origin of Y from Y_t to $Y_t + m$ (for the GDP example, from the first quarter of 1991 to the first quarter of 2019). Now, if Y_t is to be stationary, the mean, variance, and auto covariances of $Y_t + m$ must be identical to the Y.

To understand clearly if time series observations are stationary, the mean, variance, and auto covariance (at different lags) remain time invariant.

If a time series is not stationary, as mentioned above, it is called a non-stationary time series. To specify, non-stationary time series will have either time varying mean or variance or both. Time series need to be stationary because if a time series is non-stationary the behaviour can be studied only for the given time period. Each set of time series data will, therefore, be for a particular time frame and, hence, cannot be generalized. Such models have little or no use for forecasting and prediction.

The stochastic process is purely random if it has zero mean, constant variance σ^2, and is serially uncorrelated (no autocorrelation). The error term u_t is assumed to be a white noise process, denoted as $u_t \sim \text{IIDN}(0, \sigma^2)$; that is, u_t is independently and identically distributed as a normal distribution with zero mean and constant variance.

13.5 NON-STATIONARY STOCHASTIC PROCESSES

The research interest lies always in stationary time series, however, researchers do encounter non-stationary time series, and the classic example is the Random Walk Model (RWM). It is usually observed that asset prices, such as stock prices or exchange rates, follow a random walk, that is, they are non-stationary. There are two types of random walks: (1) random walk without drift (i.e., no constant or intercept term) and (2) random walk with drift (i.e., intercept is present).

13.5.1 Random Walk without Drift

Suppose that u_t is a white noise error term with mean 0 and variance σ^2. Then the time series Y_t is said to be a random walk if:

$$Y_t = Y_{t-1} + u_t \tag{13.4}$$

In the random walk model, as above in Equation 13.4, the value of Y at time t is equal to its value at time (t – 1) plus a random shock; hence, considered as AR (1) model mentioned earlier. Assume (Equation 13.4) as a regression of Y at time t on its value lagged by one period. For example, economists propagating the efficient capital market hypothesis believe that stock prices are essentially random and, hence, market cannot speculate profitability.

Now, from Equation 13.4:

$Y_1 = Y_0 + u_1$

$Y_2 = Y_1 + u_2 = Y_0 + u_1 + u_2$

$Y_3 = Y_2 + u_3 = Y_0 + u_1 + u_2 + u_3$

In general, if the process started at some time 0 with a value of Y_0, we have:

$$Y_t = Y_0 + \sum u_t \quad (13.5)$$

Therefore,

$$E(Y_t) = E(Y_0 + \sum u_t) = \sum Y_0 \quad (13.6)$$

And

$$\text{var}(Y_t) = t\sigma^2$$

The above equations explain that the mean of Y is equal to its initial value, which is constant, but as t increases, its variance increases indefinitely, thus violating a condition of stationary. RWM without drift is a non-stationary stochastic process.

An exciting feature of RWM is the tenacity of random shocks (i.e., random errors), as explained by Equation 13.6 in which Y_t is the sum of initial Y_0 plus the sum of random shocks. Resultantly, the effect of a particular shock stays. For example, if $u_2 = 2$ rather than $u_2 = 0$, then all Y_t s from Y_2 onward will be 2 units higher and the effect of this shock never dies out.

Hence, it is said that random walk has an infinite memory. However, while Y_t is non-stationary the first difference of Y_t is stationary.

13.5.2 Random Walk with Drift

Modifying Equation 13.4,

$$Y_t = \delta + Y_{t-1} + u_t \quad (13.7)$$

Where δ is known as the drift parameter.

The name drift comes from the fact that writing preceding equation as:

$$Y_t - Y_{t-1} = \Sigma\, Y_t = \delta + u_t \quad (13.8)$$

Shows that Y_t drifts upward or downward, depending on δ being positive or negative.

Following the procedure discussed for RWM without drift, it is observed that for the RWM with drift model:

$$E(Y_t) = Y_0 + t\cdot\delta$$

$$\text{var}(Y_t) = t\sigma^2$$

In RWM with drift the mean as well as the variance increases over time, hence, violating assumption of stationary. In short, RWM, with or without drift, is a non-stationary stochastic process. It is seen that RWM is an example of what literature explains as unit root process.

NOTES

To look at the random walk with and without drift, see the two graphs below:

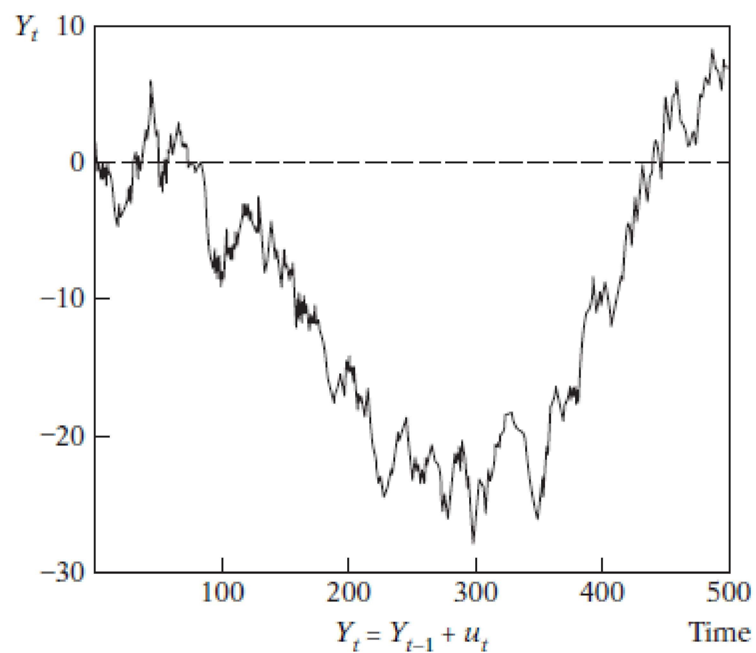


Fig. 13.1 *Random Walk without Drift*

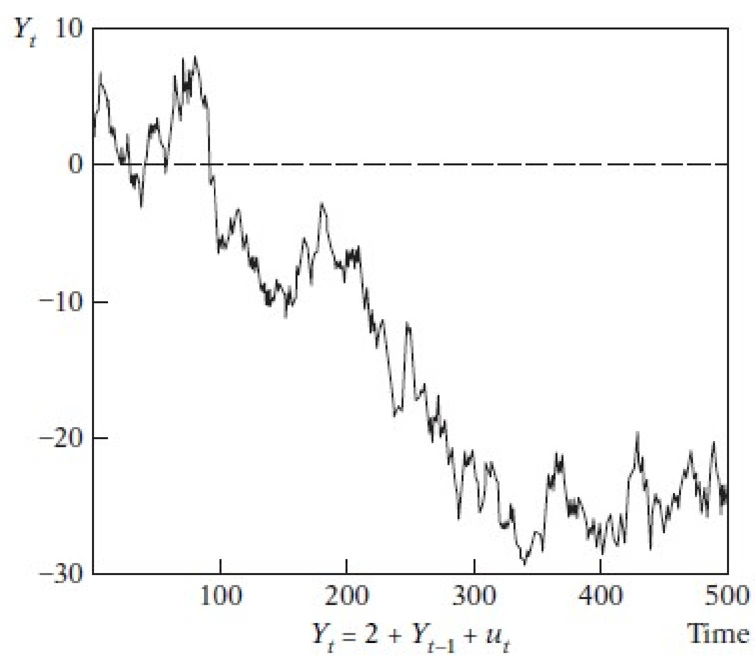


Fig. 13.2 *Random Walk with Drift*

13.6 UNIT ROOT SCHOLASTIC PROCESS

A unit root (also called a unit root process or a difference stationary process) is a stochastic trend in a time series, sometimes called a 'Random Walk with Drift'. If a time series has a unit root, it shows a systematic pattern that is unpredictable. A possible unit root.

The reason why it's called a unit root is because of the mathematics behind the process. At a basic level, a process can be written as a series of monomials (expressions with a single term). Each monomial corresponds to a root. If one of these roots is equal to 1, then that's a unit root.

Unit root tests are tests for stationary in a time series. A time series has stationary if a shift in time does not cause a change in the shape of the distribution; unit roots are one cause for non-stationary.

These tests are known for having low statistical power. Many tests exist, in part, because none stand out as having the most power. Tests include the following:

- The Dickey Fuller Test (sometimes called a Dickey Pantula test), which is based on linear regression. Serial correlation can be an issue, in which case the Augmented Dickey-Fuller (ADF) test can be used. The ADF handles bigger, more complex models. It does have the downside of a fairly high Type I error rate.
- The Elliott-Rothenberg-Stock Test, which has two subtypes:
  i. The *P*-test takes the error term's serial correlation into account.
  ii. The DF-GLS test can be applied to detrended data without intercept.
- The Schmidt-Phillips Test includes the coefficients of the deterministic variables in the null and alternate hypotheses. Its two subtypes are:
  i. The rho-test
  ii. The tau-test
- The Phillips-Perron (PP) Test is a modification of the Dickey Fuller test, and corrects for autocorrelation and heteroscedasticity in the errors.
- The Zivot-Andrews Test allows a break at an unknown point in the intercept or linear trend.

In probability theory and statistics, a unit root is a feature of some stochastic processes (such as, random walks) that can cause problems in statistical inference involving time series models. A linear stochastic process has a unit root if 1 is a root of the characteristic equation of the process. Such a process is non-stationary but does not always have a trend.

If the other roots of the characteristic equation lie inside the unit circle, that is, have a modulus (absolute value) less than one, then the first difference of the process will be stationary; otherwise, the process will need to be differenced multiple

NOTES

times to become stationary. If there are *d* unit roots, the process will have to be differenced *d* times in order to make it stationary. Due to this characteristic, unit root processes are also called difference stationary.

Let us write the RWM $Y_t = Y_{t-1} + u_t$ as:

$$Y_t = \rho\, Y_{t-1} + u_{t-1} \leq \rho \leq 1 \tag{13.9}$$

This model resembles the Markov first-order autoregressive model explained under autocorrelation. If $\rho = 1$, (Equation 13.9) becomes a RWM (without drift). If ρ is 1, there is a problem known as the unit root problem, that is, a situation of non-stationarity; here the variance of Y_t is not stationary. The name unit root is due to the fact that $\rho = 1$. In practice, it is crucial to find out if a time series possesses a unit root. Below are several tests of unit root, that is, several tests of stationarity.

13.7 THE UNIT ROOT TEST

A test of stationarity (or non-stationarity) which has gained importance in several years is the unit root test. The starting point is the unit root (stochastic) process as discussed above.

$$Y_t = \rho\, Y_{t-1} + u_{t-1} \leq \rho \leq 1 \qquad \text{(as in Equation 13.9)}$$

Where u_t is a white noise error term.

If in the above equation, $\rho = 1$, that is, in the case of the unit root, becomes a random walk model without drift, which we know is a non-stationary stochastic process. Hence, one can simply regress Y_t on its (one period) lagged value Y_{t-1} and find out if the estimated ρ is statistically equal to 1. If it is, then Y_t is non-stationary. This is the general idea behind the unit root test of stationarity.

For theoretical reasons, manipulating (Equation13.9) as follows:

Subtract Y_{t-1} from both sides of (Equation 13.9) to obtain:

$$Y_t - Y_{t-1} = \rho\, Y_{t-1} - Y_{t-1} + u_t$$

$$= (\rho - 1)\, Y_{t-1} + u_t$$

Which can be alternatively written as:

$$\Delta\, Y_t = \delta\, Y_{t-1} + u_t \tag{13.10}$$

Where $\delta = (\rho - 1)$ and Δ, as usual, is the first-difference operator. Therefore, in practice, instead of estimating (Equation 13.9), estimate (Equation 13.10) and test the (null) hypothesis that $\delta = 0$.

If $\delta = 0$, then $\rho = 1$, that is, there is a problem of a unit root, and the time series under examination is non-stationary.

While estimating Equation 13.10, it may be noted that if $\delta = 0$, then (Equation 13.10) will become:

$$\Delta\, Y_t = (Y_t - Y_{t-1}) = u_t \tag{13.11}$$

Since u_t is a white noise error term, it is stationary, which means that the first differences of a random walk time series are stationary. For estimating Equation 13.10, take the first differences of Y_t and regress them on Y_{t-1} and observe the estimated slope coefficient (= $\hat{\delta}$) = 0 or not. If it is zero, it can be clearly said that Y_t is non-stationary. But if it is negative, Y_t is stationary.

Let us look at which test should one conduct for finding out if the estimated coefficient of Y_{t-1} in (Equation 13.11) is zero or not. Regrettably, we cannot use *t*-test as with the null hypothesis $\delta = 0$ (i.e., $\rho = 1$). The *t* value of the estimated coefficient of Y_{t-1} does not follow the t distribution even in large samples because does not have an asymptotic normal distribution.

To provide the solution Dickey and Fuller have shown that under the null hypothesis that $\delta = 0$, the estimated t value of the coefficient of Y_{t-1} in (equation 13.11) follows the τ (tau) statistic.

These authors have computed the critical values of the tau statistic on the basis of Monte Carlo simulations.

The tables are now incorporated in several econometric packages. As per convention, the tau statistic or test is known as the Dickey–Fuller (DF) test, in the name of its discoverers.

However, not to forget to mention that if the hypothesis that $\delta = 0$ is rejected (i.e., the time series is stationary), (Student's) t test can be used. The actual procedure of implementing the DF test involves several decisions. In discussing the nature of the unit root process it was observed that a random walk process may have no drift, or it may have drift, or it may have both deterministic and stochastic trends.

To accommodate the various possibilities, the DF test is estimated using three different null hypotheses.

Y_t is a random walk: $\Delta Y_t = \delta Y_{t-1} + u_t$ (13.12)

Y_t is a random walk with drift: $\Delta Y_t = \beta_1 + \delta Y_{t-1} + u_t$ (13.13)

Y_t is a random walk with drift

Around a stochastic trend: $\Delta Y_t = \beta_1 + \beta_{2t} + \delta Y_{t-1} + u_t$ (13.14)

Where t is the time or trend variable.

In each case, the null hypothesis is that:

$\delta = 0$; that is, there is a unit root and the time series is non-stationary.

The alternative hypothesis is that δ is less than zero; that is, the time series is stationary.

Rejecting the null hypothesis means that Y_t is a stationary time series with zero mean in Equation 13.12 and that Y_t is stationary with a nonzero mean [$= \beta_1/(1 - \rho)$] in Equation 13.13 and that Y_t is stationary around a deterministic trend in Equation 13.14.

NOTES

Crucial to note that the critical values of the tau test to test the hypothesis that $\delta = 0$, are different for each of the preceding three specifications of the DF test.

Further, if specification (Equation 13.14) is correct, but we estimate (Equation 13.13), we will be committing a specification error.

The actual estimation procedure is as follows:

Estimate (Equation 13.12), or (Equation 13.13), or (Equation 13.14) by OLS; divide the estimated coefficient of Y_{t-1} in each case by its standard error and calculate (τ) tau statistic; and refer to the DF tables of any statistical package. If the tau computed absolute value exceeds the DF or MacKinnon critical tau values, we reject the hypothesis that $\delta = 0$, proving that time series is stationary. On the other hand, if the computed $|\tau|$ does not exceed the critical tau value, it proves the time series is non-stationary and the null hypothesis is not to be rejected.

13.8 INTEGRATED SCHOLASTIC PROCESSES

The RWM is a specific case of a general stochastic model known as integrated processes. As mentioned earlier, RWM without drift is non-stationary. However, first difference of the same is stationary. Hence, RWM without drift is termed integrated of order 1, represented by $I(1)$. Likewise, if a time series has to be differenced twice (means taking the first difference of the first differences) to make it stationary, such a time series is termed integrated of order d. Broadly, when a non-stationary time series is differenced d times to make it stationary, such time series observations are said to be integrated of order d. A time series Y_t integrated of order d is denoted as $Y_t \sim I(d)$.

If any time series like Y_t is stationary right from the beginning not requiring any differencing, it is said to be integrated of order zero, denoted by $Y_t \sim I(0)$. Hence, the 'Stationary Time Series' can also be specified as 'Time Series Integrated of Order Zero'. Majority of the economic time series are usually of the integration $I(1)$, meaning that such time series become stationary only after taking their first differences.

Features of Integrated Series

The following properties of integrated time series may be noted.

Assume X_t, Y_t, and Z_t be three time series:

1. If $X_t \sim I(0)$ and $Y_t \sim I(1)$, then $Z_t = (X_t + Y_t) = I(1)$

   This states that the sum of stationary and non-stationary time series is also non-stationary.

2. If $X_t \sim I(d)$, then $Z_t = (a + b X_t) = I(d)$, where a and b are constants.

   This states that the linear combination of an I(d) series is also I(d). Thus, if $X_t \sim I(0)$, then $Z_t = (a + b X_t) \sim I(0)$.

3. If Xt ~ $I(d_1)$ and $Y_t \sim I(d_2)$, then $Z_t = (a X_t + b Y_t) \sim I(d_2)$, where $d_1 < d_2$.
4. If $X_t \sim I(d)$ and $Y_t \sim I(d)$, then $Z_t = (a X_t + b Y_t) \sim I(d^*)$; d* is generally equal to d.

NOTES

13.9 UNDERSTANDING SPURIOUS REGRESSION

To understand the importance of stationarity, consider the following two RWMs.

$$Y_t = Y_{t-1} + u_t \quad (13.15)$$

$$X_t = X_{t-1} + v_t \quad (13.16)$$

Assume that 500 observations of u_t from $u_t \sim N(0, 1)$ and 500 observations of v_t from $v_t \sim N(0, 1)$ are generated keeping the initial values of both Y and X as zero. Also, assume that u_t and v_t are serially and mutually uncorrelated. Understand that both these time series are non-stationary; that is, they are $I(1)$ or show stochastic trends. Processes, the R_2 from the regression of Y on X should have value = 0; that is, no relationship between the two variables. Nevertheless, look at the below regression results:

Table13.1 *Regression Results*

Variable	Coefficient	Std. error	t statistic
C	−13.2556	0.6203	−21.36856
X	0.3376	0.0443	7.61223
	R^2 = 0.1044	d = 0.0121	

Observe that the coefficient of X is highly statistically significant, and the R_2 value is low, it is statistically significant and not = 0. These results tend to conclude that there exists a significant statistical relationship between Y and X, whereas theoretically there must be any. This kind of regression results are known as phenomenon of spurious or as nonsense regression, first discovered by Yule. Yule showed that nonsense or spurious correlation could persist in non-stationary time series even when the sample size is sufficiently big. The above regression is not correct as suggested by the extremely low Durbin-Watson d value. The DW test suggests strong first-order autocorrelation. As a good rule of thumb if $R_2 > d$ be suspicious that the estimated regression is spurious and the results obtained are meaningless.

Test of Stationarity

A common assumption in many time series techniques is that the data are stationary. A stationary process has the property that the mean, variance and autocorrelation structure do not change over time. Stationarity can be defined in precise mathematical terms, but for our purpose we mean a flat looking series, without trend, constant variance over time, a constant autocorrelation structure over time and no periodic fluctuations (seasonality). For practical purposes, stationarity can usually be determined from a run sequence plot.

NOTES

Transformations to Achieve Stationarity

If the time series is not stationary, we can often transform it to stationarity with one of the following techniques:

- We can difference the data. That is, given the series Z_t, we create the new series

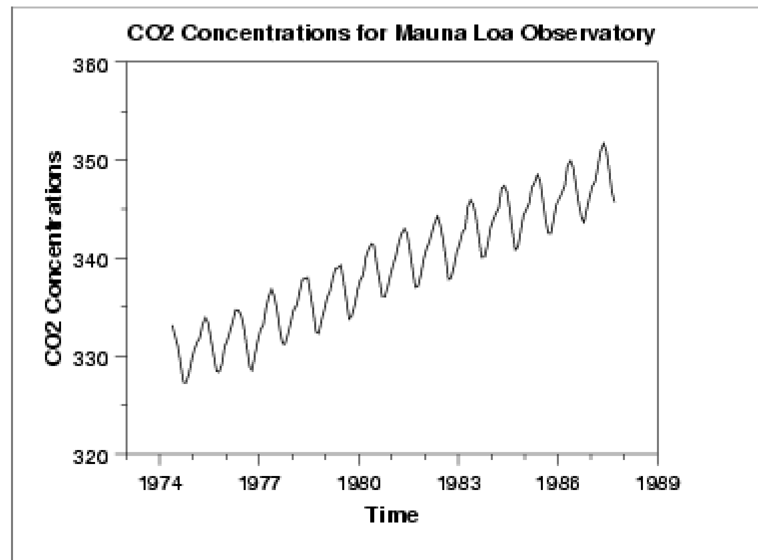
  $Y_i=Z_i-Z_i-1$

  The differenced data will contain one less point than the original data. Although you can difference the data more than once, one difference is usually sufficient.
- If the data contain a trend, we can fit some type of curve to the data and then model the residuals from that fit. The purpose of the fit being to simply remove long term trend, a simple fit, such as a straight line, is typically used.
- For non-constant variance, taking the logarithm or square root of the series may stabilize the variance. For negative data, you can add a suitable constant to make all the data positive before applying the transformation. This constant can then be subtracted from the model to obtain predicted (i.e., the fitted) values and forecasts for future points.

The above techniques are intended to generate series with constant location and scale. Although seasonality also violates stationarity, this is usually explicitly incorporated into the time series model.

Example 13.1 The following plots are from a data set of monthly CO_2 concentrations.

Solution: Run sequence plot



The initial run sequence plot of the data indicates a rising trend. A visual inspection of this plot indicates that a simple linear fit should be sufficient to remove this upward trend.

There are two practical challenges:

(1) To find out if a given time series is stationary or not

(2) If time series is not stationary, is there a way that it can be made stationary?

The two test that are prominently discussed in the literature for finding out whether times series is stationary or not are:

(1) Graphical analysis

(2) Autocorrelation function Correlogram test

1. Graphical Analysis

As a basic rule of econometric analysis before one pursues formal tests, it is always advisable to plot the time series under study. The plot gives an initial clue about the expected nature of the time series.

2. AutoCorrelation Function (ACF) and Correlogram

The more formal and yet simple test of stationarity is based on the AutoCorrelation Function (ACF).

The ACF at lag k, denoted by ρk, is defined as:

$\rho k = \gamma_k/\gamma_0$ = covariance at lag k / variance

Where covariance at lag k and variance are:

Variance: var $(Y_t) = E\,(Y_t - \mu)2 = \sigma^2$

Covariance: $\gamma k = \text{E}\,[(Y_t - \mu)(Y_{t+k} - \mu)]$

Note that if $k = 0$, $\rho 0 = 1$.

Both covariance and variance are measured in the same units of measurement; ρk is a unit less, or pure, number. Its value lies between –1 and +1, as any correlation coefficient does. If we plot ρk against k, the graph we obtain is known as the population correlogram.

N reality there is only a realization (i.e., sample) of a stochastic process, which helps in computing a sample autocorrelation function (SAFC), $\rho\hat{}k$. To calculate this, first compute the sample covariance at lag k, $\gamma\hat{}k$, and the sample variance, $\gamma\hat{}_0$, which are defined as:

$\gamma\hat{}\text{k} = \Sigma\,(Y_t - {}^{-}\text{Y})(\,Y_{t+k} - {}^{-}\text{Y}) / n$

$\gamma\hat{}0 = \Sigma\,(Y_t - {}^{-}\text{Y})2 / n$

Where, n is the sample size and ⁻Y is the sample mean.

Hence, the sample autocorrelation function at lag k is:

$\rho\hat{}k = \gamma\hat{}k / \gamma\hat{}_0$

Which is simply the ratio of sample covariance (at lag k) to sample variance. A plot of $\rho\hat{}k$ against k is known as the sample correlogram.

NOTES

Assume that a sample of 500 error terms is generated and their correlogram is presented in below two figures.

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	–0.022	–0.022	0.2335	0.629
		2	–0.019	–0.020	0.4247	0.809
		3	–0.009	–0.010	0.4640	0.927
		4	–0.031	–0.031	0.9372	0.919
		5	–0.070	–0.072	3.4186	0.636
		6	–0.008	–0.013	3.4493	0.751
		7	0.048	0.045	4.6411	0.704
		8	–0.069	–0.070	7.0385	0.532
		9	0.022	0.017	7.2956	0.606
		10	–0.004	–0.011	7.3059	0.696
		11	0.024	0.025	7.6102	0.748
		12	0.024	0.027	7.8993	0.793
		13	0.026	0.021	8.2502	0.827
		14	–0.047	–0.046	9.3726	0.806
		15	–0.037	–0.030	10.074	0.815
		16	–0.026	–0.031	10.429	0.843
		17	–0.029	–0.024	10.865	0.863
		18	–0.043	–0.050	11.807	0.857
		19	0.038	0.028	12.575	0.860
		20	0.099	0.093	17.739	0.605
		21	0.001	0.007	17.739	0.665
		22	0.065	0.060	19.923	0.588
		23	0.053	0.055	21.404	0.556
		24	–0.017	–0.004	21.553	0.606
		25	–0.024	–0.005	21.850	0.644
		26	–0.008	–0.008	21.885	0.695
		27	–0.036	–0.027	22.587	0.707
		28	0.053	0.072	24.068	0.678

AC = autocorrelation, PAC = partial autocorrelation Q-Stat = Q statistic, Prob = probability.

Observe the column labelled AC, which is the sample autocorrelation function, and the first diagram on the left, labelled autocorrelation. The solid vertical line in this diagram represents the zero axis. Observations above the line are positive values and those below the line are negative values. As is very clear from this diagram, for a purely white noise process the autocorrelations at various lags revolves around zero. This is the picture of a correlogram of a stationary time series. Therefore, if the correlogram of an actual (economic) time series resembles the correlogram of a white noise time series, we can say that time series is probably stationary.

For the choice of length of lag, a rule of thumb is to calculate ACF up to one-third to one-quarter the length of the time series used for the analysis.

Further, in order to avoid the spurious regression problem which result from regressing a non-stationary time series? Therefore, it is important to transform non-stationary time series to make them stationary. The process includes:

Difference Stationary Process (DSP): In case a time series has a unit root, the first difference of such time series are stationary. Hence, the solution here is to take the first differences of the time series.

Trend Stationary Process (TSP): Trend stationary process is a process where time series is stationary around the trend line. In order to make such times series stationary is to regress it on time and the residuals from the regression will then be stationary.

If the time series is DSP and treated as TSP, it is termed under-differencing. On the other hand, if the time series is TSP but treated as DSP then it is referred to as over-differencing.

Check Your Progress

1. Explain the analysis of economic time series.
2. Define the stochastic process.
3. State the stationary stochastic processes.
4. Illustrate the non-stationary stochastic processes.
5. Define the unit root scholastic process.
6. Elaborate on the unit root test.
7. Interpret the integrated scholastic processes.

13.10 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The times series data is defined as data values of the variables at equally spaced time interval. The analysis takes into account that the data observations taken over the time period have an internal structure, such as autocorrelation, specific trends or having seasonal variations.
2. A random or stochastic process is an assembly of random variables which are ordered in time. Let Y be a random variable, and if it is continuous then denoted it as $Y(t)$, but if it is discrete, denoted it as Y_t.
3. Generally, a stochastic process is considered to be stationary if its mean and variance are constant over time and the value of the covariance between the two time periods depends only on the distance or gap or lag between the two time periods and not the actual time at which the covariance is computed.

NOTES

4. The research interest lies always in stationary time series, however, researchers do encounter non-stationary time series, and the classic example is the Random Walk Model (RWM). It is usually observed that asset prices, such as stock prices or exchange rates, follow a random walk, that is, they are non-stationary.
5. A unit root (also called a unit root process or a difference stationary process) is a stochastic trend in a time series, sometimes called a 'Random Walk with Drift'. If a time series has a unit root, it shows a systematic pattern that is unpredictable. A possible unit root.
6. A test of stationarity (or non-stationarity) which has gained importance in several years is the unit root test. The starting point is the unit root (stochastic) process,

   $Y_t = \rho\ Y_{t-1} + u_{t-1} \leq \rho \leq 1$

   Where u_t is a white noise error term.
7. When a non-stationary time series is differenced d times to make it stationary, such time series observations are said to be integrated of order d. A time series Y_t integrated of order d is denoted as $Y_t \sim I(d)$.

13.11 SUMMARY

- The times series data is defined as data values of the variables at equally spaced time interval. The analysis takes into account that the data observations taken over the time period have an internal structure, such as autocorrelation, specific trends or having seasonal variations.
- The key objective of time series analysis is to develop an understanding of the underlying phenomenon and structure represented by the observed data. Further, the idea is to fit a model that leads to forecasting, monitoring and feedback, and forward control.
- A random or stochastic process is an assembly of random variables which are ordered in time. Let Y be a random variable, and if it is continuous then denoted it as $Y(t)$, but if it is discrete, denoted it as Y_t.
- GDP is a stochastic variable and to understand the same assume the GDP of $2872.8 billion for 1991–I quarter. According to theory, the GDP value for the first quarter of 1991 could have been any random number, depending on the economic and political environment prevailing in the nation.
- Generally, a stochastic process is considered to be stationary if its mean and variance are constant over time and the value of the covariance between the two time periods depends only on the distance or gap or lag between the two time periods and not the actual time at which the covariance is computed.

- If a time series is not stationary, as mentioned above, it is called a non-stationary time series. To specify, non-stationary time series will have either time varying mean or variance or both.
- The stochastic process is purely random if it has zero mean, constant variance σ^2, and is serially uncorrelated (no autocorrelation).
- The research interest lies always in stationary time series, however, researchers do encounter non-stationary time series, and the classic example is the Random Walk Model (RWM).
- A unit root (also called a unit root process or a difference stationary process) is a stochastic trend in a time series, sometimes called a 'Random Walk with Drift'. If a time series has a unit root, it shows a systematic pattern that is unpredictable. A possible unit root.
- Unit root tests are tests for stationary in a time series. A time series has stationary if a shift in time does not cause a change in the shape of the distribution; unit roots are one cause for non-stationary.
- In probability theory and statistics, a unit root is a feature of some stochastic processes (such as, random walks) that can cause problems in statistical inference involving time series models.
- A linear stochastic process has a unit root if 1 is a root of the characteristic equation of the process. Such a process is non-stationary but does not always have a trend.
- Broadly, when a non-stationary time series is differenced *d* times to make it stationary, such time series observations are said to be integrated of order d. A time series Y_t integrated of order d is denoted as $Y_t \sim I(d)$.
- A common assumption in many time series techniques is that the data are stationary. A stationary process has the property that the mean, variance and autocorrelation structure do not change over time.

13.12 KEY WORDS

- **Time series data:** It is defined as data values of the variables at equally spaced time interval.
- **Time series analysis:** The analysis takes into account that the data observations taken over the time period have an internal structure, such as autocorrelation, specific trends or having seasonal variations.
- **Stochastic processes:** A random or stochastic process is an assembly of random variables which are ordered in time. Let *Y* be a random variable, and if it is continuous then denoted it as $Y(t)$, but if it is discrete, denoted it as Y_t.

- **Stochastic stationary process:** Generally, a stochastic process is considered to be stationary if its mean and variance are constant over time and the value of the covariance between the two time periods depends only on the distance or gap or lag between the two time periods and not the actual time at which the covariance is computed.
- **Unit root scholastic process:** A unit root (also called a unit root process or a difference stationary process) is a stochastic trend in a time series, sometimes called a 'Random Walk with Drift'.
- **Unit root test:** A test of stationarity (or non-stationarity) which has gained importance in several years is the unit root test.
- **Integrated scholastic processes:** When a non-stationary time series is differenced *d* times to make it stationary, such time series observations are said to be integrated of order d. A time series Y_t integrated of order d is denoted as $Y_t \sim I(d)$.
- **Test of stationarity:** A stationary process has the property that the mean, variance and autocorrelation structure do not change over time.

13.13 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What is analysis of economic time series?
2. Explain the stochastic process.
3. Define the stationary stochastic processes.
4. Interpret the non-stationary stochastic processes.
5. State the unit root scholastic process.
6. What do you understand by the unit root test?
7. Explain the integrated scholastic processes.

Long-Answer Questions

1. Briefly discuss the analysis of economic time series.
2. Explain the stochastic process with the help of examples.
3. Differentiate between the stationary stochastic processes and non-stationary stochastic processes.
4. Describe the two types of random walks in non-stationary stochastic processes.
5. Explain the unit root scholastic process. Write the importance of unit root test.

6. Analyse the integrated scholastic processes. Give appropriate examples.
7. What is test of stationarity? Explain the transformations to achieve stationarity.

13.14 FURTHER READINGS

Johnston, J. and John DiNARDO. 1997. *Econometric Methods*, Fourth Edition. New Delhi: Tata McGraw-Hill.

Koutsoyiannis, A.1977. *Theory of Econometrics,* Second Edition. London: The Macmillan Press Ltd.

Özdemir, Durmuº. 2016. *Applied Statistics for Economics and Business,* Second Edition. Izmir (Turkey): Springer.

Maddala, G. S. 1992. *Introduction to Econometrics,* Second Edition. New York: Macmillan Publishing Company.

Pindyck, R. S and D. L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts,* Fourth Edition. New York: McGraw Hill.

Goldberger, A. S. 1998. *Introductory Econometrics***.** Cambridge: Harvard University Press.

Levine, David M., Timothy C. Krehbiei, Mark L. Berenson and P. K. Viswanathan. 2009. *Business Statistics*, Fifth Edition. New Delhi: Pearson Education.

Webster, Allen L. 1998. *Applied Statistics for Business and Economics,* Third Edition. New Delhi: Tata McGraw-Hill.

UNIT 14 ECONOMETRIC SOFTWARE PACKAGE: STATA

Structure

14.0 INTRODUCTION

The statistical software STATA is easy to use. It is extremely optimized for common econometric methods, for both computation and syntax. It has a large user base which is active and is of great support to each other as a community.

When STATA's program icon is used to start STATA by double-clicking it, STATA's interface opens up. At the top of the screen in this interface there are various top-down menus and short-cut buttons that invoke various commands, such as, for browsing/editing data, saving a file and performing statistical analysis. Of these menus, the Help menu can be used to obtain guidance for using the various STATA commands as well as for accessing information from the built-in reference manual. The Help menu also enables users to lookup updates, as also connect directly connect official website of STATA. Most of the rest of the available GUI screen of STATA is divided between the four windows also called panes/panels/work area.

Learning and using STATA can be done by using one of two different approaches. Let us look at them both. One approach is to look at it as being an interactive tool and use it as such. To do this, just invoke STATA, load the data into it and get into clicking or typing commands. This provides a great way for data exploration, and for using various commands to check how they implement, check

out syntax, and see how the program works. This is a great way to learn stuff in STATA as it will show immediate results which you can check to see if they are what you required and tweak your commands if it led to different results. You can also access immediate Help from the software for whatever you wish to do with the software. Yet, it must be kept in mind that interactive work cannot be easily or reliably reproduced, or modified. Mistakes cannot be set right as STATA does not provide any command to undo a command that has been executed.

NOTES

The second approach to learning and working with STATA is to consider it to be a programming language. When this approach is taken, a complete program is written, called do files, and they are then run. The do file comprises those very commands that would be entered one by one at the STATA command line, and because in the program these are written in a permanent file, it is possible to catch the problems with the commands, correct them and run the file again. Since they are in the form of a program, they contain the trail of how the result was arrived at. They are a record of the path followed for obtaining a specific result. For those who are looking to rely on any results, present results, or even to publish them, going the program way is the right option for those projects.

In this unit, you will study about the econometric software package: STATA.

14.1 OBJECTIVES

After going through this unit, you will be able to:

- Define what econometric software package: STATA is
- Explain the work area of the STATA interface
- Analyse the STATA commands for performing various tasks

14.2 INTRODUCTION TO STATA

The complex and trivial econometric analysis are now carried out with ease with the use of various statistical software packages like SPSS, E Views, R, JAMOVI, and STATA. Generally, those packages that are simple to use come with minimal features and analytical tools. Various packages available are complex to handle and require coding knowledge as well. One such statistical analysis software is STATA which has vast comprehensive features and is also very simple to operate. With STATA, majority of the analysis can be carried out by using its simple drop down menus and the reading of the result it provides and its interpretation is quite easy.

STATA is a powerful statistical software that enables users to analyse, manage, and produce easy graphs, histograms and charts for data visualization. It is predominantly used by researchers in the fields of economics, biomedicine, and political science, to examine data patterns and analyse the same for inferential

NOTES

analysis. It has both a command line and graphical user interface making the software more intuitive and easy to use.

To begin with, when you open STATA and enter the software, the screen will look like that depicted in Figure 14.1. Notice that the screen is divided into 4 panels or areas.

Area A is called the command line. This is the area where one will type statements which will be required to analyse data or simply give a command. Area B displays the variable list when the data set is loaded into STATA; all the variables chosen for performing the task specified in the command line will be listed in the box. Area C is known as review box and this contains a history of all the commands used during the particular STATA session. Area D is where the generated results are reported.

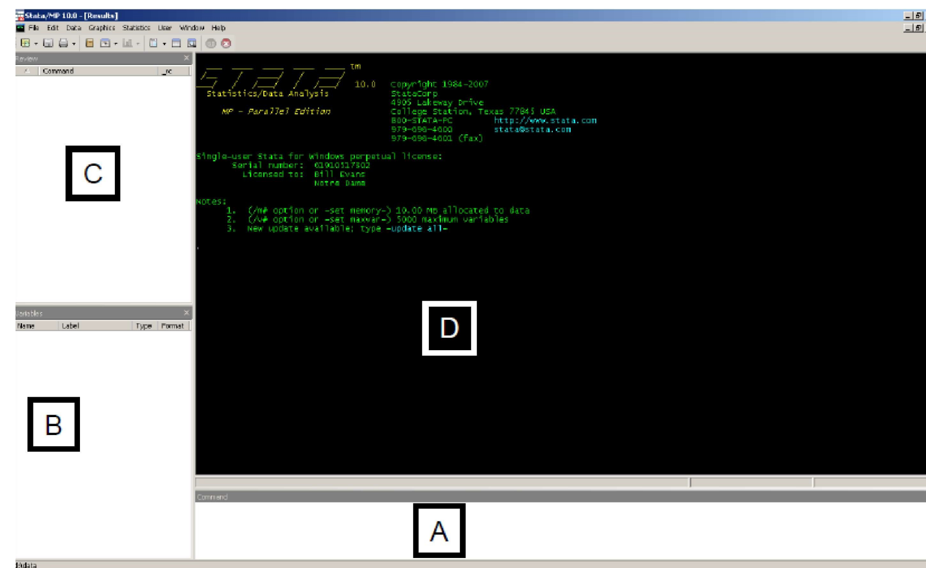


Fig. 14.1 *Command Line in STATA*

Command line is the bar on the screen available for typing all the necessary commands. The contents of the other boxes depend on the command bar information. When STATA is invoked, the cursor at the command line will begin to blink, indicating that the system is ready to accept input. To execute a command, type it in and pressing the Enter key.

In this unit, anything mentioned in COURIER FONT is a command that is executable by using the command line. STATA provides for two different ways to generate statistics. One way is by writing executable statements, line by line at the command line, and running the codes. Alternatively, you write an entire program that has a group of executable statements, and submit this as a single whole from the command line.

NOTES

The Command bar can be used to obtain help with syntax and usage of commands. Suppose that a user needs information about how to describe the contents of data sets. At the command line, just type:

```
help describe
```

Now, press the Enter key and a pop-up box appears with the syntax for the 'describe' command. It is interesting to note that the executed command is now in the Review box C hence at any time it is possible to re-use a command that has already been executed. To do this, using the mouse, click once on the command and the statement re-appears in the command line.

STATA considers that all external files are stored on the default subdirectory (folder) depending upon how a particular machine is set up. The user can create a subdirectory for STATA work, and using STATA, change the default folder. Here is an example of how to do it. Presume that a folder d:\bill\econ30331 is created for storing all STATA related work. In the command bar type:

```
cd d:\bill\econ30331
```

And press the enter key. Now, STATA will search in this folder for all data sets and also store all results to this folder. While working interactively, the user must save a 'Log' of the command activity and results from the working STATA session that are posted in the results section (area D in Figure 14.1).

To create a log, type the command:

```
log using stata_log_1.log,replace
```

And press the enter key. The log will be written to the file 'stata_log_1.log' and the replace option tells the program to overwrite a current file with that name. At the end type:

```
log close
```

And press the enter key to close the log file. Please note that STATA commands, data set names, and variable names are case sensitive.

14.3 OPENING A STATA DATA FILE

A data set comprises a collection of variables that describe different units to be used for analysis. Assume that the data set is as a matrix of columns and rows. The rows are separate observations (individual, firms, cities, time periods) while each column is a different variable that describes a specific characteristic of the observations in the sample used for the study. For many analyses, there is access to a STATA data file that is already in STATA format and ready for use by the program. STATA data files possess a .dta extension and loading them into STATA is simple and straight.

STATA is a very fast working program that requires that all data be read into RAM for quick processing. Therefore, the constraint on the program is usually

NOTES

the available RAM. The program will not let the data set load in case the available RAM is insufficient. RAM allocation is dependent on the machine being used. However, one may allocate more RAM to STATA at any time during a STATA session. For most of the class assignments, 2 meg of RAM should be sufficient.

Set RAM by typing in the command bar:

```
set memory 2m
```

And press the enter key.

Assume the data set available as cps80.dta (it is possible to choose any data file and even name it as preferred).

To load the data into STATA, type:

```
use cps80
```

And press the Enter key. The data variables associated with the specified data set are now available for use in STATA.

New variables can be constructed with the dataset in memory, one can delete particular observations, and also generate statistics. After constructing new variables, save the revised data set by using the command:

```
save cps80_update
```

Or change the old name by using the command:

```
save cps80, replace
```

To clear the memory of a data set no longer in use, type:

```
clear
```

And press the enter key.

14.4 READING RAW DATA INTO STATA

For most empirical assignments, read the data into a STATA data file. This can be accomplished through a variety of different steps. A simple method is to transport data from a spread sheet like EXCEL into a STATA data set. Figure 14.2 below displays the first 32 lines from an EXCEL data set cps80.xls. The data file is a matrix with 7 columns and 19,906 rows. Each row depicts data for another observation (individual) and the column observation is a new variable. Assume that this data file consists of males, aged 21-64 who worked full time (>30 hours per week) during any survey. The variables in order are: age, race, years_educ, union_status, smsa_size, region, and weekly_earn. Observe detailed description in Figure 14.3 below and understanding by the variable definitions in Figure 14.3,

the first observation is for a 55 year old white man with 12 years of education, in a union, from one of the largest 19 standard metropolitan statistical areas in the northeast and making ₹ 750 per week.

NOTES

EXCEL stores data differently than STATA so it is required to transform it to a format called 'comma delimited' data, or CSV format. In CSV format, each row is stored on a different line and the variables are separated by a comma.

Opening the data set into a program editor the data appears as shown in Figure 14.3. It is important to observe that the first row contains variable names while the other rows have the data points. All rows are on different lines and all variables are separated by commas. Now, the data is ready for STATA to read in the current format into a STATA data file. At the command line, if you type:

```
insheet using cps80.csv, comma
```

Then, press the enter key, the data will be loaded into STATA.

The results box indicates that 7 variables and 19,906 observations were loaded up into STATA.

It is good programming practice to LABEL all variables with a short description of the variables as this will prove useful at a later stage. To provide a label for the variable age, type:

```
Label var age "age in years"
```

Then press the enter key. Sample labels for the other 6 variables are listed below.

```
Label var race "=1 if white non-Hisp, =2 if black non-Hisp, =3 if Hispanic"

Label var years_educ "years of competed education"

Label var union_status "=1 if in union, =2 otherwise"

Label var smsa_size "=1 if largest 19 smsa, =2 if other smsa, =3 not in smsa"

Label var region "=1 if northeast, =2 if midwest, =3 if south, =4 if west"

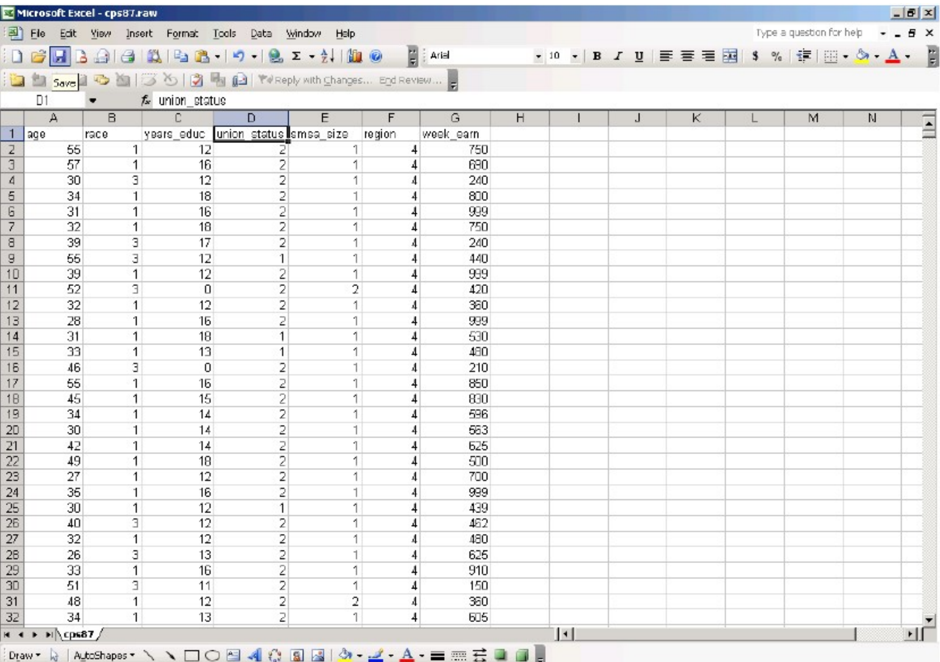
Label var weekly_earn "usual weekly earnings, up to $999"
```

At any time, a list of all of the variables can be generated by typing:

```
describe
```

NOTES

And pressing the enter key. A description of the data set is mentioned as Block A after giving the data command.



	age	race	years_educ	union_status	smsa_size	region	week_earn
2	55	1	12	2	1	4	750
3	57	1	16	2	1	4	690
4	30	3	12	2	1	4	240
5	34	1	18	2	1	4	800
6	31	1	16	2	1	4	999
7	32	1	18	2	1	4	750
8	39	3	17	2	1	4	240
9	56	3	12	1	1	4	440
10	39	1	12	2	1	4	999
11	52	3	0	2	2	4	420
12	32	1	12	2	1	4	360
13	28	1	16	2	1	4	999
14	31	1	18	1	1	4	530
15	33	1	13	1	1	4	480
16	46	3	0	2	1	4	210
17	55	1	16	2	1	4	850
18	45	1	15	2	1	4	830
19	34	1	14	2	1	4	596
20	30	1	14	2	1	4	563
21	42	1	14	2	1	4	625
22	49	1	18	2	1	4	500
23	27	1	12	2	1	4	700
24	35	1	16	2	1	4	999
25	30	1	12	1	1	4	439
26	40	3	12	2	1	4	462
27	32	1	12	2	1	4	480
28	26	3	13	2	1	4	625
29	33	1	16	2	1	4	910
30	51	3	11	2	1	4	150
31	48	1	12	2	2	4	360
32	34	1	13	2	1	4	605

***Fig. 14.2** Content of Data Set CPS 80*

Variable	Definition
AGE	Age in years
RACE	=1 if white, non-Hispanic, =2 if black, non-Hispanic, =3 if Hispanic
EDUC	Years of completed education, maximum is 18.
UNIONM	=1 if a union member, =2 otherwise
SMSA	=1 if live in one of 19 largest Standard metropolitan Statistical Areas (SMSA), =2 if live in other SMSA, =3 if live in non-SMSA
REGION	=1 if live in Northeast, =2 if live in Midwest, =3 if live in South, =4 if live in West
EARNWKE	Usual weekly earnings, nominal 1987 dollars, maximum is $999

***Fig. 14.3** Contents of cps80.csv*

```
age,race,years_educ,union_status,smsa_size,region,week_earn
55,1,12,2,1,4,750
57,1,16,2,1,4,690
30,3,12,2,1,4,240
34,1,18,2,1,4,800
31,1,16,2,1,4,999
```

NOTES

```
32,1,18,2,1,4,750
39,3,17,2,1,4,240
55,3,12,1,1,4,440
39,1,12,2,1,4,999
52,3,0,2,2,4,420
32,1,12,2,1,4,360
28,1,16,2,1,4,999
31,1,18,1,1,4,530
33,1,13,1,1,4,480
46,3,0,2,1,4,210
55,1,16,2,1,4,850
45,1,15,2,1,4,830
34,1,14,2,1,4,596
30,1,14,2,1,4,563
42,1,14,2,1,4,625
49,1,18,2,1,4,500
27,1,12,2,1,4,700
35,1,16,2,1,4,999
30,1,12,1,1,4,439
40,3,12,2,1,4,462
```

Block A cps80.do

```
* set the memory to 2 meg
set memory 2m
* set it such that the computer does not
* need the operator to hit the return key
* to continue
set more off
* write results to a log file
log using cps80.log,replace
* read in raw data from comma delimited data
insheet using cps80.csv, comma
* label the variables
Label var age "age in years"
Label var race "=1 if white non-Hisp, =2 if black non-Hisp, =3 if Hispanic"
```

NOTES

```
    Label var years_educ "years of competed
education"
    Label var union_status "=1 if in union, =2
otherwise"
    Label var smsa_size "=1 if largest 19 smsa, =2
if other smsa, =3 not in smsa"
    Label var region "=1 if northeast, =2 if midwest,
=3 if south, =4 if west"
    Label var weekly_earn "usual weekly earnings,
up to $999"* describe what is in the
    data set
    describe
    * generate new variables
    * lines 1-2 illustrate basic math functoins
    * line 3 line illustrates a logical operator
    * line 4 illustrate the OR statement
    * line 5 illustrates the AND statement
    gen age2=age*age
    genln_weekly_earn=ln(weekly_earn)
    gen union=union_status==1
    gen nonwhite=((race==2)|(race==3))
    gen big_ne=((region==1)&(smsa==1))
    label var age2 "age squared"
    label var ln_weekly_earn "log earnings per week"
    label var union "1=in union, 0 otherwise"
    label var nonwhite "1=nonwhite, 0=white"
    label var big_ne "1= live in big smsa from
northeast, 0=otherwsie"
    * get descriptive statistics for all variables
    sum
    * get statistics for only a subset of variables
    sum age years_educ
    * get detailed descriptics for a subset of
variables
    sumweekly_earn age, detail
```

```
    * to get means across different subgroups in the
    * sample, first sort the data, then generate
    * summary statistics by subgroup
    sort race
    by race: sum weekly_earn
    * get weekly earnings for only those with a
    * high school education
    Sum weekly_earn if years_educ>=12
    * get frequencies of discrete variables
    tabulate race
    * get two-way table of frequencies
    tabulate region smsa, row column
    * test whether means are the same across two subsamples
    Ttest weekly_earn, by(union)
    *run simple regression
    regln_weekly_earn age age2 years_educnonwhite union
    * run regression adding smsa, region and race fixed-effects
    xi: regln_weekly_earn age age2 years_educ union i.racei.regioni.smsa
    * close log file
    log close
```

NOTES

14.5 DEVELOPING NEW VARIABLES IN STATA

STATA provides the 'gen' command for generating new variables. The syntax for the 'gen' command is:

`Gen` *`new variable name=mathematic expression`*

The new variable is the name of the generated variable and it must follow STATA naming conventions. The basic rules for naming variables are:

- STATA is case-sensitive.
- Names can contain no more than 32 characters.
- Variables can contain letters, numbers, or underscores (_).

NOTES

- Spaces or other special characters (such as &,* and%) are not allowed.
- The first character cannot be a number. It can be a letter or underscore.

Consider the example given below. These will generate new variables from that data set that was earlier in the unit loaded into STATA.

```
gen age2=age*age
gen ln_weekly_earn=ln(weekly_earn)
gen union=union_status==1
gen nonwhite=((race==2)|(race==3));
gen big_northeast_city=((region==1)&(smsa==1));
```

One of the most common variables in applied work is a dummy variable that takes the value 1 or 0, separating individual into two groups (male or female, black or white, etc). These variables are easy to construct by using 'Logical Operators'.

Logical operators are of the form gen y=(logical statement) that construct a new variable Y that equals 1 when the logical statement is true and zero otherwise.

The last three variables generated above, demonstrate how to use logical operators. The variable union constructs a variable that equals 1 for union members and zero otherwise. Notice that two equal signs must be used when exact equality is indicated in a logical statement. Combinations of logical statements can be used to construct dummy variables. The vertical line '|' represents 'OR' and the '&' sign represent 'AND'.

After the variables are constructed, add a set of variable LABELs. The command for labels is shown below:

```
Label var age2 "age squared"
Label var ln_weekly_earn "ln usual earnings per week"
Label var union "1=in union, 0 otherwise"
Label var nonwhite "1=nonwhite, 0=white"
Label var big_ne "1= live in big smsa from northeast,
0=otherwsie"
```

Getting Descriptive Statistics

After data is properly loaded to STATA, one can begin with generating descriptive statistics, also called summary statistics, (mean, min, max and standard deviation) by using the command:

```
sum
```

NOTES

It provides descriptive statistics for all variables. For specific information for a subset of variables, like age and education, add the variables after the sum command as shown below:

```
sum age years_educ
```

And press the Enter key. For more detailed information on a particular variable (quantiles, medians, skewness, kurtosis, etc.), use the 'sum' command and list the variables.

```
sum weekly_earn age, detail
```

The above command generates detailed statistics for only two variables. Results from these three exercises are reported in blocks B, C and D respectively.

Block B: Results

cps80.log

```
________________________________________

log: d:\bill\stata\cps80.log
log type: text
opened on: 12 Aug 2008, 12:22:05
.
. * read in raw data from comma delimited data
. insheet using cps87.csv, comma
(7 vars, 19906 obs)
.
. * label the variables
. label var age "age in years"
. label var race "=1 if white non-Hisp, =2 if
black non-Hisp, =3 if Hispanic"
. label var years_educ "years of competed
education"
. label var union_status "=1 if in union, =2
otherwise"
. label var smsa_size "=1 if largest 19 smsa,
=2 if other smsa, =3 not in smsa"
. label var region "=1 if northeast, =2 if
midwest, =3 if south, =4 if west"
. label var weekly_earn "usual weekly earnings,
up to $999"
. * describe what is in the data set
```

NOTES

```
. describe
Contains data
obs: 19,906
vars: 7
size: 318,496 (88.6% of memory free)
```

Box A

```
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
age             byte   %8.0g                  age in years
race            byte   %8.0g                  =1 if white non-Hisp, =2 if
                                                black non-Hisp, =3 if Hispanic
years_educ      byte   %8.0g                  years of competed education
union_status    byte   %8.0g                  =1 if in union, =2 otherwise
smsa_size       byte   %8.0g                  =1 if largest 19 smsa, =2 if
                                                other smsa, =3 not in smsa
region          byte   %8.0g                  =1 if northeast, =2 if midwest,
                                                =3 if south, =4 if west
weekly_earn     int    %8.0g                  usual weekly earnings, up to
                                                $999
-------------------------------------------------------------------------------
```

Fig. 14.4 *Box A*

Box B

```
. * get descriptive statistics for all variables
. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |     19906    37.96619    11.15348         21         64
        race |     19906    1.199136     .525493          1          3
  years_educ |     19906    13.16126    2.795234          0         18
union_status |     19906    1.769065    .4214418          1          2
   smsa_size |     19906    1.908369    .7955814          1          3
-------------+--------------------------------------------------------
      region |     19906    2.462373    1.079514          1          4
 weekly_earn |     19906     488.264    236.4713         60        999
        age2 |     19906    1565.826    912.4383        441       4096
ln_weekly_~n |     19906    6.067307     .513047   4.094345   6.906755
       union |     19906    .2309354    .4214418          0          1
-------------+--------------------------------------------------------
    nonwhite |     19906    .1408118    .3478361          0          1
      big_ne |     19906    .1409625    .3479916          0          1
```

Fig. 14.5 *Box B*

Box C

```
. * get statistics for only a subset of variables
. sum age years_educ

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |     19906    37.96619    11.15348         21         64
  years_educ |     19906    13.16126    2.795234          0         18
```

Fig. 14.6 *Box C*

Box D

```
. * get detailed descriptics for a subset of variables
. sum weekly_earn age, detail

              usual weekly earnings, up to $999
-------------------------------------------------------------
      Percentiles      Smallest
```

Fig. 14.7 *Box D*

NOTES

In Block B, observe that the mean age is 37.97 years and 23% of workers are in unions. In Box D, observe that median weekly earnings are $449 dollars but average earnings are higher at $488.26.

To look at average weekly earnings across different racial and ethnic groups, first sort the data.

```
sort race
```

Then, run a command for means calculated for the racial subgroups

```
by race: sum weekly_earn
```

The *by variable*: Option must be ended with a colon (:). The *by* option can be used with virtually all of STATA's commands.

To generate complete distributions for discrete variables, use the TABULATE command. For example, if you wanted to know the fraction of people by racial/ethnic group, you would type:

```
tabulate race
```

And press the enter key.

Box E

```
. * get frequencies of discrete variables
. tabulate race

=1 if white |
  non-Hisp, |
=2 if black |
  non-Hisp, |
     =3 if  |
   Hispanic |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |     17,103       85.92       85.92
          2 |      1,642        8.25       94.17
          3 |      1,161        5.83      100.00
------------+-----------------------------------
      Total |     19,906      100.00
```

Fig. 14.8 *Box E*

Box E gives the results for the above mentioned tabulate command. It shows that 85.9 percent of the sample is white, non- Hispanic, 8.25 are Black, non-Hispanic while 5.83% are Hispanic.

14.5.1 Testing Whether Means in Two Subsamples are the Same

The simplest statistical test that can be conducted is to study whether the means from two different groups are the same. Let's take an example of examining weekly earnings for union and non-unions workers. The difference in means across samples is tested with a t-test and the STATA command is

```
Ttest weekly_earn, by(union)
```

The results from this exercise are reported in Box F. Observe that the mean earnings among unions workers is $515.28 while the mean earnings for non-union workers is $480.15 and therefore, the difference across the two groups (non-union

NOTES

minus union) is -$35.13. The *t*-statistic calculated for the difference is -27.35. The 95% critical value of a t-test with 19,904 degrees of freedom is 1.96 and so we reject the null hypothesis that the means across the two subsamples are the same, also indicated by low *p* value.

Box F

```
. * test whether means are the same across two subsamples
. ttest weekly_earn, by(union)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |   15309    480.1503    2.017734    249.6532    476.1953    484.1053
       1 |    4597    515.2845    2.705061    183.4063    509.9813    520.5878
---------+--------------------------------------------------------------------
combined |   19906     488.264    1.676048    236.4713    484.9788    491.5492
---------+--------------------------------------------------------------------
    diff |           -35.13423    3.969334               -42.91446     -27.354
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -8.8514
Ho: diff = 0                                     degrees of freedom =    19904

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Fig. 14.9 *Box F*

14.5.2 Running a Simple OLS Regression

One of the most significant tools in STATA is to run a simple OLS regression. Simple regressions are generated by the reg command and the syntax is basic where the first variable after reg is the dependent variable and all other variables are independent variables. Let the example have five covariates: age, age2, years_educ, union and non-white. STATA on its own adds a constant to every model unless otherwise specified. The regression syntax used is as follows.

```
reg ln_weekly_earn age age2 years_educ nonwhite union
```

The regression results for a chosen example are reported in BOX G.

BOX G

```
. *run simple regression
. reg ln_weekly_earn age age2 years_educ nonwhite union

      Source |       SS       df       MS              Number of obs =   19906
-------------+------------------------------           F(  5, 19900) = 1775.70
       Model |  1616.39963     5  323.279927           Prob > F      =  0.0000
    Residual |  3622.93905 19900  .182057239           R-squared     =  0.3085
-------------+------------------------------           Adj R-squared =  0.3083
       Total |  5239.33869 19905  .263217216           Root MSE      =  .42668

------------------------------------------------------------------------------
ln_weekly_~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0679808   .0020033    33.93   0.000     .0640542    .0719075
        age2 |  -.0006778   .0000245   -27.69   0.000    -.0007258   -.0006299
  years_educ |    .069219   .0011256    61.50   0.000     .0670127    .0714252
    nonwhite |  -.1716133   .0089118   -19.26   0.000    -.1890812   -.1541453
       union |   .1301547   .0072923    17.85   0.000     .1158612    .1444481
       _cons |   3.630805   .0394126    92.12   0.000     3.553553    3.708057
------------------------------------------------------------------------------
```

Fig. 14.10 *Box G*

14.5.3 Clearing and Closing of the Analysis

After the interactive STATA session is completed, the log file can be closed by typing:

```
log close
```

NOTES

And pressing the Enter key. Also, in order to exit, clear the data out of memory of STATA by typing:

```
Clear
```

And pressing the enter key.

The data is cleared out of memory at this point.

Check Your Progress

1. Give the name of three statistical software packages.
2. What is the area in the user interface of STATA known as where a user can input a statement and execute by pressing the Enter key?
3. Which command in STATA will enable the user obtain the syntax for how to describe the contents of data sets?
4. Which command will be used to close a log file?
5. Explain the file extension used by STATA data files.
6. Define the connection between STATA being a fast working program and the RAM.
7. State the command which will be used to assign 3 meg of RAM for STATA.
8. Which command will be used to clear the memory of a data set no longer in use?
9. Illustrate the purpose of the command: describe.

14.6 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Some of the statistical software packages are SPSS, E Views, R, JAMOVI, and STATA.
2. The area in the user interface of STATA where a user can input a statement and execute by pressing the Enter key is called the command line.
3. `help describe`
4. `log close`
5. The file extension used by STATA data files is .dat.
6. STATA is a very fast working program that requires that all data be read into RAM for quick processing.

NOTES

7. `set memory 3m`
8. clear
9. At any time, one can get a list of all of the variables in the data set by typing: describe.

14.7 SUMMARY

- Complex and trivial econometric analysis are now carried out with ease with the use of various statistical software packages like SPSS, E Views, R, JAMOVI, and STATA. With STATA, majority of the analysis can be carried out by using its simple drop down menus and the result reading and interpretation it is quite easy.
- STATA is a powerful statistical software that enables users to analyse and manage data, and produce easy graphs, histograms and charts for data visualization.
- STATA is predominantly used by researchers in the fields of economics, biomedicine, and political science, to examine data patterns and analyse the same for inferential analysis.
- STATA has a command line as well as a graphical user interface that make the software more intuitive and easy to use.
- The opening screen of STATA is split into 4 panels/areas. Area A is the command line where commands are entered or statements which will be required to analyse data are typed. Area B displays the variable list chosen for performing the task when a data set is loaded into STATA. Area C is the review box which contains a history of all the commands used during a particular STATA session. Area D is where the generated results are reported.
- STATA provides for two different ways to generate statistics. One, by writing executable statements, line by line at the command line, and running the codes. Alternatively, by creating an entire program that has a group of executable statements, and submitting this as a single whole from the command line.
- A pop-up box appears with the syntax for the command. It is interesting to note that the executed command is now in the Review box C hence at any time it is possible to re-use a command that has already been executed.
- It is possible to change the default working directory directly from the STATA software. For example: cd d:\bill\econ30331.
- Here is an example of creating a log in STATA: log using stata_log_1.log, replace.

NOTES

- For most empirical assignments, read the data into a STATA data file. This can be accomplished through a variety of different steps. A simple method is to transport data from a spread sheet like EXCEL into a STATA data set.
- EXCEL stores data differently than STATA so it is required to transform it to a format called 'comma delimited' data, or CSV format. In CSV format, each row is stored on a different line and the variables are separated by a comma. Here is an example of loading such data into STATA: insheet using cps80.csv, comma.
- It is good programming practice to LABEL all variables with a short description of the variables which will prove useful at a later stage. Here is an example of providing a label for the variable age: labelvar age "age in years".
- At any time, a list of all of the variables can be generated by typing: describe.
- Use the 'gen' command for generating new variables. The syntax for the 'gen' command is: Gen *new variable name=mathematic expression.*
- One of the most common variables in applied work is a dummy variable that takes the value 1 or 0, separating individual into two groups (male or female, black or white, etc). These variables are easy to construct by using 'logical operators'.
- To generate complete distributions for discrete variables, use the TABULATE command.
- The simplest statistical test than can be conducted is to study whether the means from two different groups are the same.
- One of the most significant tools in STATA is to run a simple OLS regression. Simple regressions are generated by the *reg* command and the syntax is basic where the first variable after *reg* is the dependent variable and all other variables are independent variables.

14.8 KEY WORDS

- **Command line:** An interface for typing commands directly to a computer's operating system.
- **Data set:** A collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer.
- **Syntax:** the system of rules for the structure of a command in a language.
- **Raw data:** Raw data is unprocessed data.
- **Case sensitive:** Requiring correct input of uppercase and lowercase letters.
- **OLS regression:** Ordinary Least Squares (OLS) regression is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable.

14.9 SELF ASSESSMENT QUESTIONS AND EXERCISES

NOTES

Short-Answer Questions

1. Elaborate on the dummy variable.
2. What is summary statistics, and what is the command used for it?
3. State the two ways in which statements and commands can be given to STATA.
4. Define the purpose and usage of insheet command.

Long-Answer Questions

1. Discuss briefly the various rules for naming variables in STATA.
2. Explain the command: `Label var race "=1 if white non-Hisp, =2 if black non-Hisp, =3 if Hispanic"`
3. How does the data appear when the data set is opened in a program editor?
4. Analyse the reg command.

14.10 FURTHER READINGS

Johnston, J. and John DiNARDO. 1997. *Econometric Methods*, Fourth Edition. New Delhi: Tata McGraw-Hill.

Koutsoyiannis, A.1977. *Theory of Econometrics,* Second Edition. London: The Macmillan Press Ltd.

Özdemir, Durmuº. 2016. *Applied Statistics for Economics and Business,* Second Edition. Izmir (Turkey): Springer.

Maddala, G. S. 1992. *Introduction to Econometrics,* Second Edition. New York: Macmillan Publishing Company.

Pindyck, R. S and D. L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts,* Fourth Edition. New York: McGraw Hill.

Goldberger, A. S. 1998. *Introductory Econometrics***.** Cambridge: Harvard University Press.

Levine, David M., Timothy C. Krehbiei, Mark L. Berenson and P. K. Viswanathan. 2009. *Business Statistics*, Fifth Edition. New Delhi: Pearson Education.

Webster, Allen L. 1998. *Applied Statistics for Business and Economics,* Third Edition. New Delhi: Tata McGraw-Hill.